Low complexity in-loop perceptual video coding

Yetish Girish Joshi

Supervisors: Dr. J. Loo

Dr. P. Shah

Dr. S. Rahman

Dr. A. Tasiran

Advisors: Mr. G. Hearne

Mr. L. Miraziz

Faculty of Science and Technology

Middlesex University

This dissertation is submitted for the degree of

Doctor of Philosophy

London October 2016

મમી અને પાપા,

પ્રેમ, વિશ્વાસ, ધીરજ

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Yetish Girish Joshi

October 2016

Acknowledgements

I would like to acknowledge my supervisors, members of my school, staff of the University and my colleagues, past and present. It was Dr Loo, who fuelled my interest in video coding and pushed my technical boundaries. Through Dr Shah I developed my critical thinking skills and was encouraged to established the quality of workmanship expected. While the feedback provided by Dr Rahman helped to structure my research for an academic audience. Equally, Dr Tasiran has aided in handling the data produced by introducing to me the R programming language. The workshops run by Dr Duncker and presented by Prof. Woodman informed the need to develop as a researcher across the different stages along the PhD journey. My change encounters at the kitchen with Prof Wong, would usually impart some perspective, especially encouraging reflection on dead-ends as experience. Following the 2^{nd} Middlesex University summer research conference, Dr Mapp, highlighted that I needed to optimise my approach. In writing, the support by way of access to booklets for my colleagues and myself from Paula (Bernaschina) helped to address the feedback given by my supervisors. While the discussions with Gary (Hearne), provided a means to redirect or reinforce my limited statistical and experimental design knowledge. Finally, Leonard (Miraziz) gave encouragement and stories of my peers to whom I could relate with and my colleagues gave me support, especially that by Pawel (Chwalinski).

I would also like to extend my acknowledgements to the technical software tools of Dia, yEd, Veusz, R, XeLaTeX and the video coding codebase from Joint Video Team (JVT). Both Dia and yEd were used to illustrate concepts and processes presented in this body of work. Veusz to produce graph figures, while R for exploratory data analysis and modelling. XeLaTeX is the typesetting system derived from LaTeX which enabled production of papers and this thesis. Finally, the codebase of H.264/AVC and HEVC from JVT which enabled the development of the proposed perceptual video coding solutions.

I can live with doubt and uncertainty and not knowing. I think it is much more interesting to live not knowing than to have answers that might be wrong. If we will only allow that, as we progress, we remain unsure, we will leave opportunities for alternatives. We will not become enthusiastic for the fact, the knowledge, the absolute truth of the day, but remain always uncertain …In order to make progress, one must leave the door to the unknown ajar.

Richard Feynman

The first principle is that you must not fool yourself, and you are the easiest person to fool.

Richard Feynman

Look at me!
Look at me!
Look at me NOW!
It is fun to have fun
But you have to know how.

Dr. Seuss

The Cat in the Hat

Abstract

The tradition of broadcast video is today complemented with user generated content, as portable devices support video coding. Similarly, computing is becoming ubiquitous, where Internet of Things (IoT) incorporate heterogeneous networks to communicate with personal and/or infrastructure devices. Irrespective, the emphasises is on bandwidth and processor efficiencies, meaning increasing the signalling options in video encoding. Consequently, assessment for pixel differences applies uniform cost to be processor efficient, in contrast the Human Visual System (HVS) has non-uniform sensitivity based upon lighting, edges and textures. Existing perceptual assessments, are natively incompatible and processor demanding, making perceptual video coding (PVC) unsuitable for these environments. This research allows existing perceptual assessment at the native level using low complexity techniques, before producing new pixel-base image quality assessments (IQAs). To mange these IQAs a framework was developed and implemented in the high efficiency video coding (HEVC) encoder. This resulted in bit-redistribution, where greater bits and smaller partitioning were allocated to perceptually significant regions. Using a HEVC optimised processor the timing increase was $< +4\%$ and $< +6\%$ for video streaming and recording applications respectively, 1/3 of an existing low complexity PVC solution. Future work should be directed towards perceptual quantisation which offers the potential for perceptual coding gain.

Contents

List of Figures

List of Tables

Nomenclature

Greek Symbols

$\sigma_{x,y}$	Covariance of original and reconstructed blocks, it indicates how related they are in magnitude and direction
σ_x	Standard deviation, σ_x, original block, σ_y, reconstructed block
σ_x^2	Variance for a block, σ_x^2, original block, σ_y^2, reconstructed block

Superscripts

C_n	Constants, C_1 and C_2, used in SSIM to stabilise equation. $C_n = (K_n L)^2$, for 8 bit Luma, K_1 is 0.01 and K_2 is 0.03

Other Symbols

$\bar{x}$	Mean for an array of pixels (block), $\bar{x}$, original block, $\bar{y}$, reconstructed block
λ	Lagrange multiplier (λ), quantisation applied to achieve desired bit-rate, as part of the RDO process
λ_p	Perceptual re-scaling of Lagrange multiplier (λ), quantisation applied to achieve desired bit-rate, in SSIM based PVC as part of the RpDO process

$\trianglelefteq$ Triangle Inequality, where the sum of two sides is greater than the third, a condition a metric must satisfy

D Distortion cost, related to a given bit-rate and quantisation, part of the RDO process

D_p Perceptual distortion assessment cost, related to a given bit-rate and quantisation, part of the RDpO process

J_{min} Minimum total energy, sum of distortion cost with Lagrange multiplier (λ) with bit-rate, used when performing RDO

R Bit-rate used along with Lagrange multiplier (λ) as part of the RDO process

Acronyms / Abbreviations

5G Fifth generation of mobile and wireless communication network, covering a variety of wireless networks interoperating across a range of frequencies

6LoWPAN IPv6 low power wireless personal area networks, used to represent portable limited resource sensors which communicate within and between heterogeneous networks

ACR Absolute category rating, where under subjective testing the participant is asked to rate the proposed video sequence against a scale without the original

AMP Asymmetric motion partitioning, part of HEVC prediction where blocks height to width ratios are 1:4, 3:4, 4:1 or 4:3

AR Augmented reality, where real-world video has 3D graphics or high resolution videos super imposed

CB Coding block, like that of a CU, but representing the respective luma, colour (chrominance) parts individually.

CGI Computer Graphics Image, synthesising/rendering images with the use of computers

CIF Common interchange format, 352x288 pixels

CSF Contrast sensitivity function, based upon a series of experiments to describe the non-linear response of the HVS to changes in lighting

CU Coding unit, the quad-tree representation of a leaf/node that encompasses both prediction and transform information of a block-size

DCR Degradation category rating, where under subjective testing the participant is asked to rate the proposed video sequence against a scale relative to the original

DVB Digital video broadcasting

DVD Digital video disc

EBU European Broadcast Union, consortium of broadcasters to promote best practise and standards

fps Frames per second

GOP Group of frames

H.264/AVC H.264 advance video coding

H.265 also known as HEVC (High efficiency video coding standard)

HD High Definition, where the resolution is greater than 768 by 576 pixels per frame up to an area of 2Mpxs

HDR High dynamic range, support for larger range of contrast, which can improve the immersive experience for the viewer

HEVC High efficiency video coding standard

HFR High frame rate, increases support for up to 120 fps

HVS Human Visual System

IoM Internet of Media, where video or images are captured and transmitted using low powered and/or remote devices which communicate using heterogeneous networks to the internet

IoT Internet of Things, where low power and/or remote devices which communicate using heterogeneous networks to the internet

IPv6 Internet protocol version 6, where up to 2^{128} ($3.4x10^{38}$) the unique number of addressable devices can be connected on the same network

IQA Image Quality Assessment, defining image quality by its retention of perceptually significant markers

ITU International Telecommunications Union

JND Just noticeable difference, an inverse of the CSF to produce threshold based detection to lighting levels

LCU Largest coded unit, simliar to macroblock, instead of being fixed to 16x16, the LCU can be 16x16, 32x32 or 64x64

LTE Long-Term Evolution

MOS Mean opinion score

MPEG Motion Picture Experts Group

Mpx Mega pixels

MSE Mean of square errors

NSS Natural Scene Statistics, where statistical moment information is used to perform IQA

PB Prediction block, represents the prediction (sub-) block level within a given coding block (CU)

PC Pairwise comparison, where under subjective testing the order of the original and proposed is randomised

PSNR Peak signal to noise ratio, traditionally used as an objective measure for image quality

PU Prediction unit, the prediction (signalling) portion of the coding unit for a given block

PVC Perceptual video coding, where perceptual techniques are applied in video coding

QCIF Quarter common interchange format, 176x144 pixels

QP Quantisation parameter, used to represent the compression applied, with a value ranging from 0 to 51, calculated from λ

RDO Rate-distortion optimisation, represented with the R-D curve, used to find the optimal between distortion and bit-rate

RoF Radio over fibre, where radio networks coverage are extended over fibre optics

ROI Region of interest, assessing what would act as perceptual clues by the HVS in order to judge a scene or recognise/track an object

SAD Sum of absolute differences

SATD Sum of absolute transform differences, which involves using the Hadamard transform on the differences

SD Standard Definition, 704x576 pixels

SDV Software defined video, where off-the-shelf hardware is used to perform video encoding driven by software, rather than specialised processors

SSE Sum of square errors

SSIM Structural Similarity, a IQA that uses NSS to provide an image quality score, used as an objective measure

STDM Standard traditional distortion metric, used as a term to classify SAD, SSE and SATD, which account for differences irrespective of the HVS sensitivity to the original

UBR Universal bounded region, the region of shared space occupied by STDM and SSIM

UHD Ultra high definition, between two to four times the horizontal and vertical resolution of HD, resulting in 8Mpx and 33.1Mpx per frame

V1 The initial stage of the HVS modelled by CSF and JND, reflecting the filtering taken place by the eye

V2 The feature map generation by the HVS used reduce the information transported across the optic nerve

V3 The stage where spatial recognition occurs

V4 The final stage where object tracking happens

VCD Video compact disc

VCEG Video Coding Experts Group, part of the ITU which standardises H.26x series of video decoders

VCL Video coding layer, used to represent the sub-block coding structure, containing also the prediction and residual information

VoD Video on Demand, where video content is streamed to devices as individually requested than broadcast simultaneously

VR Virtual reality, where immersive 3D graphics or high resolution 360 degrees videos are presented typically with the use of a headset

WCG Wide colour gamut, which extends supported coverage of colour from 35.9% to 75.8% by moving from ITU Rec. 709 to Rec. 2020

WSN Wireless sensor networks, where low power and/or remote devices communicate using heterogeneous networks

.

Chapter 1

Introduction

Video is a powerful medium, and the technology that encapsulates it is equally important as the message it delivers. As such video-based services are increasingly popular across the Internet, with video-based data estimated to represent over three quarters of all the Internet traffic (Cisco, 2015). In terms of global mobile data traffic, over half is video-related content and this is expected to increase as 4th generation (4G) Long-Term Evolution (LTE) networks and handsets are deployed (Cisco, 2015). Irrespective, demand and limited bandwidth availability to deliver video is leading to radio over fibre solutions (RoF), where heterogeneous networks such as fibre backbone and Wi-Fi/4G at the last stage are applied. Similarly, the Internet is being used as a platform by content creators and a movement for software defined video (SDV) is allowing video coding to be technology agnostic. This movement towards Internet based video is also being spurred by the demand for video where spectrum bandwidth is limited. Consequently, traditional broadcasters where high definition (HD) has matured, are discussing to establish an Ultra-HD (UHD) standard, which is twice the horizontal and vertical resolution of HD.

To support UHD and compatibility with low power devices across heterogeneous networks, the high efficiency video coding (HEVC) standard was produced, offering compression rates of up to 50% compared to its predecessor of H.264 advance video codec (H.264/AVC). The HEVC codec is designed in much the same way as its predecessor H.264/AVC of exploiting redundancy with more options. This is recognised with projects such as PROVISION, *Per*ceptually *O*ptimised *Vi*deo Compres*sion* which are evaluating video means to further exploit perceptual related redundancy in video coding (Provision-itn.eu, 2015). Perceptual evaluation relates to incorporating models based upon the human visual system (HVS) so that content may be compressed more intelligently (Lee and Ebrahimi, 2012). These HVS models are adopted in perceptual based video coding as an alternative means to identify redundancy, and thus, perceptual video coding (PVC) presents a new avenue to extend existing standards in bandwidth limiting environments.

The use of a PVC solution could also be used in upcoming standards, where there is debate about the next video coding broadcast standard. Within the European Broadcast Union (EBU), discussions are for an UHD standard called UHD-1 Phase 2. UHD-1 Phase 2 supports BT.2020, which includes high frame rate (HFR), where frame rates may go up to 120 frames per second (fps), as well as offering higher dynamic range (HDR) and wide colour gamut (WCG) which refers to 10 or 12 bit depth for Luma and colour respectively (Borer and Cotton, 2015; ITU.int, 2015; Noland, 2014). This does increase the amount of data to be encoded, however, due to statistical redundancy the average bit rate difference between 60fps and 120fps is 8% (Gabriellini, 2014). This means that a perceptual video coding (PVC) solution has a potential role in improving bandwidth efficiency whilst maintaining the immersive user experiences.

Employing a PVC solution could perceptually re-allocate bits and led to bit savings, however, PVC introduces additional complexity which risks being a burden on resources and the overall encoding timing. As content creation is occurring on Internet enabled and portable devices, production is becoming less studio centric and more distributed leading to mobile/portable IP studio (R&D, 2012). Under such circumstances, the complexity of a video codec can be the limiting factor, for which, HEVC is known to be highly complex compared to its predecessor, due to

the additional signalling modes (Bossen et al., 2012). For this reason the issues of complexity associated for video transmitting and streaming, is recognised, leading to projects like THIRA to find complexity savings (Thira, 2015). This presents the dilemma, that while the HVS is complex to model accurately as a whole, even applying these limited HVS models as part of an PVC solution consumes substantial processing resources. However, video encoding among low powered and where computing capability is limited, such as the internet of things (IoT) is dependent upon efficient design using hardware acceleration (Smith, 2016). This need for low power friendly design is important as applications for video within IoT are considered part of the internet of media (IoM) and include multi-sensor environments of IPv6 low power wireless personal area networks (6LoWPAN) (E. G. Pereira and R. Pereira, 2015). Therefore, while video is processor intensive, hardware acceleration is being made available, however, perceptual models used in PVC solutions are highly complex liable to consume substantial power.

With the existing PVC solutions, the label of low complexity may be relative, as existing PVC solutions can be measured across the encoding of a video sequence, than by design. This is because these PVC solutions are operating outside of the native sub-block level of the hybrid-block based encoder and operating at block, frame or even sequence levels which limit the effects of PVC. As a consequence these existing PVC solutions reflect how their implementation is complexity limited and apply perceptual techniques retrospectively. Since their design is not processor friendly, they are unable to assess candidates individually. The main obstacle is that the HVS has a non-linear sensitivity response, this means existing perceptual models are highly complex by design, meaning that reducing their update frequency will reduce the overall processing load. Under these circumstances, for existing PVC solutions where update frequency is lowered when selecting candidates and/or re-scaling quantisation there is a risk that the reference used is inaccurate. Therefore, an in-loop solution is required that can run natively during the candidate selection process, however, it must also be a low complexity design. All this can be summarised in Figure 1.1, with each stage indicating the increased level of integration required to make PVC a viable solution for low powered devices.

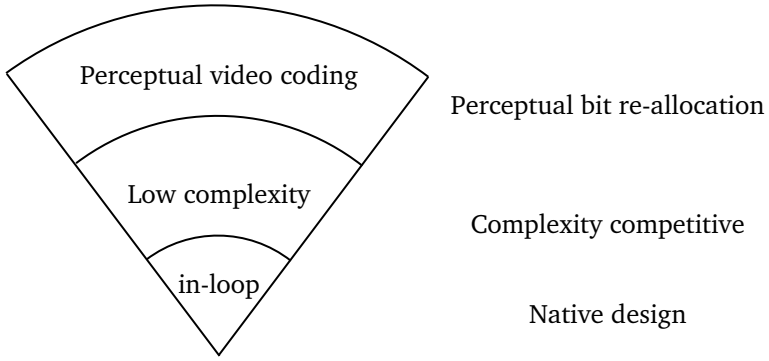


Figure 1.1 Requirements to make a PVC solution suitable for low powered devices

1.1 Research challenges

In order to address perceptual video coding at the sub-block level, existing forms of perceptual assessment, by way of an image quality assessment (IQA) must be evaluated. The most widely recognised and applied IQA is Structural Similarity (SSIM), which has been adopted within existing forms of PVC. These existing PVC solutions are attractive as they achieve perceptual quantisation and obtain a coding gain over the reference encoder. However, this improve bandwidth utilisation involves using highly complex mathematical operations, which make them unattractive for low powered devices. This is because when using perceptual related assessment or models like SSIM, scores must be rescaled to allow compatibility with existing assessments. The rescaling of perceptual scores is important as it allows the existing encoding processes to remain unchanged, which is designed along the Lagrange multiplier, used in finding the optimal point between two opposing resources (Everett III, 1963). In this case, the rescaled perceptual score must be measured against the bit usage score for each candidate, indicating that the rescaling process is a costly overhead especially as it involves non-linear operations. For this reason existing PVC solutions present the rescaling as a two stage process, first to crease a score compatible perceptual response curve at the sequence or frame level, then secondly use this curve locally to re-map existing scores.

Therefore, the challenge is to make a single process, where assessment scores are compatible and using low complexity techniques, that enables each candidate to be assessed individually. This means creating IQAs which have native support

with each existing forms of assessment, whilst using processor friendly operations to keep the complexity envelope low. However, to aid this development, an understanding of existing assessment is required, before developing new IQAs and related tools. This means when evaluating existing assessment at the sub-block level before developing new IQAs, including producing new tools to visualise the effects of IQAs on the coding structure.

1.2 Motivation

The above research challenges state that PVC should occur at the native sub-block level in a low complexity solution. This has the potential to extend PVC for applications in low powered devices. This is important as existing PVC solutions are either resource heavy or/and are non-native, making them academic as they lack a complexity competitive solution (Chandler, 2013). There is rising use of portable devices which offer personal video communications and monitoring for online and offline respectively. Conversely, user expectation for portable devices are for immersive user experiences as TVs, where portable devices are used as a main screen to commuters, second screens to TVs and provide access to a virtual/augmented (mixed) reality. This means that video content is viewed on both TVs and portable screens with similar expectation of video quality. Low complexity in-loop PVC has a role to ensure this expectation is maintained, whilst providing the encoder opportunities to improve bandwidth efficiency. Working at the native sub-block level, PVC can evaluate candidates individually, resulting in choices that re-allocate bits to retain the perceptual integrity. Therefore, in situations where the bandwidth restrictions may lower user experience, a sub-block level PVC can minimise the risk of disrupting the user experience. This is made more exciting as the latest generation of video encoder, HEVC is expected to be challenged by an open and royalty-free codec especially as open technologies for video communications are being established (Open Media, 2016; WebRTC, 2016). However, underlying these codecs is the hybrid block-based encoder principal and the need to perform distortion assessment.

1.3 Research question and objectives

Video content is being produced increasingly on low powered devices, which is driven by the popularity and accessibility to internet enabled infrastructure. Video encoding is possible among these devices as both they have processors where video encoding is hardware accelerated and support by the software to be processor friendly (Smith, 2016). This means in-loop distortion and activity assessments which accounts for substantial complexity cost is hardware accelerated or written to be processor friendly. Unfortunately, these forms of assessment are not HVS friendly, while perceptual assessment by way of an IQA is not hardware accelerated, which makes PVC unattractive on low powered devices. Since existing forms of assessment operate in-loop and at the pixel level, this means that new IQAs need to be developed that harness hardware acceleration and optimisation. This would allow PVC to be viable on low powered devices, which potentially could redistribute bits to retain the perceptual integrity of the video sequence. Therefore, the research question is presented as:

"How to make a low complexity in-loop PVC solution?"

with the following objectives:

1. Examine existing PVC solutions, appraising their technical and design features.
2. Investigate and integrate an existing IQA of SSIM into an in-loop PVC solution
3. Produce IQAs with native sub-block support which are of low complexity.
4. Implement and test the proposed low complexity in-loop PVC solution.

1.4 Structure of this thesis

These objectives reflect the research stages from the understanding of SSIM at the unexplored sub-block level, through to producing a new framework designed to overcome the issues of complexity and compatibility. Overall, these will be expressed as contribution chapters following the initial background and critique of Chapter 2 and Chapter 3 respectively. Among existing PVCs, SSIM is the most popular IQA, therefore, SSIM at the native sub-block level is investigated within

Chapter 4. This highlights the relationship between SSIM and with standard traditional distortion metrics (STDMs), where STDMs reflect where uniform cost applied to pixel differences. These STDMs consists of sum of absolute difference (SAD), sum of absolute transform difference (SATD) and sum of square errors (SSE). The observed relationship between SSIM and STDM at the sub-block level is modelled to produce a low complexity means of scaling for SSIM, to enable individual perceptual assessment of prediction candidates. While this proves perceptual at sub-block is possible, SSIM is highly complex algorithm compared to existing STDMs. Consequently, a hybrid STDM-IQA framework is proposed, along with pixel-based IQAs where perceptual assessment cost is applied when the distortion is perceptually significant. The design, implementation and testing are presented across three chapters in Chapters 6 to 8. Then Chapter 9 will evaluate whether the proposed hybrid STDM-IQA framework design goals are reflected in the findings, and how this compares to other existing work. Finally, this research is summarised as a whole in Chapter 10.

1.5 Contributions

The findings for this research have been presented in three conferences across four papers (Joshi, Loo, Shah, Rahman and Chang, 2013; Joshi, Loo, Shah, Rahman and Tasiran, 2015; Joshi, Shah et al., 2013; Xplore, 2016). These papers are listed below and are attached to the appendix in Chapter A.

1. “Review of Standard Traditional Distortion Metrics and a need for Perceptual Distortion Metric at a (Sub) Macroblock Level” in the 2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) at Brunel University, London.

2. “A novel low complexity Local Hybrid Pseudo-SSIM-SATD distortion metric towards perceptual rate control” in the 2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) at Brunel University, London.

3. “Low complexity sub-block perceptual distortion assessment for mode decision and rate-control” in the 2015 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) at Ghent, Belgium.

4. “Native in-loop prediction perceptual video coding using pixel-based IQA for HEVC” in the 2016 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) at Nara, Japan.

Chapter 2

Background of video coding, PVC and SSIM

Image and video lossy encoding encourages reusing encoded regions in order to minimise the amount stored as approximation of differences. This process is known to be effective and efficient, substantially reducing the bandwidth and storage requirements of video sequences. When this is applied to both within and between frames, operating with a pixel array block size it is known as a hybrid-block based encoder. This type of encoder forms the basis of modern video encoding standards, including MPEG 2, H.264/AVC and HEVC, which have been standardised by the International Telecommunications Union (ITU) and Motion Picture Experts Group (MPEG). As video coding standards have evolved, they require less bandwidth for the same resolution. This is possible by offering more candidates in terms of signalling and sub-block choices per block, which encourages block matching through reuse whilst minimising pixel differences. However, as the volume of candidates increases so does the overall encoding time, since all these candidates must be assessed. This means that the role of a distortion assessment becomes more prominent as video coding standards evolve. Unfortunately, existing

forms of distortion assessment are simplistic and widely regarded as a poor measure of image representation (T. N. Pappas et al., 2013; Z. Wang et al., 2004; H. R. Wu, Reibman et al., 2013).

Existing distortion metrics, known as STDMs provide a uniform cost for differences, which while they do not reflect the HVS, are computationally efficient, and remain in use. In comparison, IQAs offer a non-linear cost based upon the HVS sensitivity; however, the complexity of IQAs make them unattractive for the video coding environment. PVC is a means to bring HVS understanding into the video coding environment, where a PVC solution builds upon the video coding environment to best preserve the perceptual cues of video sequences. This is beneficial to applications of broadcast and video on demand (VoD) services, including personal video communication, tele-medicine and visual realisation.

A popular perceptual assessment that is used in PVC solutions is SSIM, an algorithm designed for image coding and considered a perceptual alternative to peak-signal to noise ratio (PSNR). SSIM operates using a sliding window approach to gather statistical measures when calculating its score, which makes it attractive for video coding. However, SSIM is incompatible with STDMs and is very complex, which hinders its adoption into PVC solutions. This chapter will first discuss the principles of video coding followed by a description of the HVS, phases of PVC solutions and finally examine the perceptual assessment of SSIM.

2.1 Video coding standards

This section is orientated around the perspective of: the historical context of the video coding standard, the underlying principles and key innovative aspects of the current generation of the video coding standard. The historical context will describe the journey to what has been a successful means for standardising the video decoder by way of efficient techniques. Then the underlying video coding principle will be examined to highlight how hybrid-block based codecs such as MPEG standard has been widely adopted. Finally, the current video coding standard of HEVC introduces some innovative means to encourage efficient use of signalling.

Standard	ITU	Year	New application	HD rate
MPEG 1	H.261	1993	Offline	N/A
MPEG II	H.262	1995	Broadcast	9Mbps
MPEG4/AVC	H.264	2003	Internet	4Mbps
HEVC	H.265	2013	Mobile	2.5Mbps

Table 2.1 Summary of video coding standards and average bit rate required for an encoded film with a PSNR of 35dB (Avsforum, 2015)

2.1.1 Brief history of the motivation of video coding standards

The standardisation of video coding over time has allowed video to become common place with increased breadth of application for each generation, as listed in Table 2.1. This was initiated with an application by Hollywood studios to lower the manufacturing cost of video tapes, resulting in MPEG 1 applied in video compact disc (VCD). While VCD was not popular due to both practical reasons of requiring multiple discs and picture quality, it did fuel the development of MPEG 2. MPEG 2 was the first to incorporate the hybrid block-based principle, used in digital versatile discs (DVDs) and for the initial digital video broadcasting (DVB) standard. In the years following MPEG 2, the application of streaming video across the Internet spurred development of H.264/AVC. H.264/AVC also formalised the move towards HD video for movie studios and broadcasters with new off-line optical media and broadcasting standards. In recent times, portable devices offer an immersive experience as they have rich multimedia functionality, enabling video communications. The HEVC standard, also known as H.265 was designed to extend support towards portable devices and encoded ultra HD (UHD) resolution videos. Overall, each generations of the video coding standard can be illustrated by its application demonstrating the proliferation of video, as shown in Figure 2.1.

The success of hybrid-block based encoders

The initial MPEG 1 standard was a proving ground for video coding standard and was restricted to a single form type of encoding, spatial coding, also known as intra coding. Therefore, except for MPEG 1, the MPEG x/H.26x series of video coding standards are hybrid-block based codecs, where the changes are represented as either spatial (intra) coding or by temporal (inter) coding techniques. Both

Optical disk (VCD)	Broadcast (DVB)	Internet streaming	Mobile network based video communication
MPEG 1	MPEG 2	MPEG 4/AVC H.264/AVC	HEVC H.265

Figure 2.1 Overview of video codecs and supported applications

coding techniques are where existing (sub-) blocks are used to identify statistical redundancies, to reduce bit usage. In intra coding, spatially adjacent (sub-) blocks based upon a set pattern are applied within a frame, while for inter, (sub-) blocks from adjacent frames which follow trajectory of movement are used between frames. By combining these different techniques significant reductions in bit usage are achieved. Each generation of video coding standards expands the intra and inter coding techniques, allowing for greater proportion of video to be represented as signalling than quantised pixels. With the increasing availability of intra and inter coding techniques the demand of processing increases too. H.264/AVC was designed to be more computationally friendly, reducing the dependency on multiplies or divides, and its success has allowed extending to portable and Internet enabled devices. With the advent of H.265 (HEVC), video is moving towards supporting UHD with higher resolutions of 4096x2160 pixels, while the underlying infrastructure requires further video coding efficiencies to meet the rising VoD services. Compared to H.264/AVC, HEVC is able to reduce the bit-rate by up to 50%, as it provides more efficient means to identify intra and inter redundancy. However, HEVC has extended inter and intra coding techniques, meaning that assessment of candidates has a more significant role.

2.2 Hybrid block-based encoder

The hybrid block-based encoder which underlies these video coding standards, is both a process to manage a video sequence into pixel array blocks, and a workflow to identify statistical redundancies. The process will be illustrated, while the workflow will be described in several stages, culminating in the balance between bit-rate and the distortion. This means that assessment will be used to find where this is optimal for distortion, while rate-control will allocate bits based on picture activity to optimise for bit-rate. As such, each of these stages exist in each implementation of a hybrid block-based encoder and are discussed in this section.

2.2.1 Video encoding process

A hybrid block-based encoder, has high compression efficiencies because video broken down to a series of group of pictures (GOP) to represent several frames together, yet each frame is broken into uniform blocks. In a simplistic view, the break down of a sequence into GOP, then into frames, blocks and sub-blocks can be illustrated in Figure 2.2. At this stage these uniform pixel arrays, known as blocks, are processed within the hybrid block-based encoder. In order to balance picture quality with respect to bit rate, the encoder considers each of these block(s) either as a series of sub-blocks or a single block, under intra or inter coding. By encoding one block at a time, the memory requirements and computational load are lowered as only a small array of pixels are considered. Furthermore, at the sub-block level use of statistical redundancy is likely to be very high, which allows sub-block reuse. Here a two stage process is undertaken by the encoder, the first is prediction at the sub-block level which leads to second, mode decision which forms the respective block. Both prediction and mode decision stages have different approaches; prediction seeks a single candidate per given sub-block size across intra and inter coding search spaces. On the other hand mode decision uses those successful candidates from prediction to find the combination of sub-blocks to represent the block. Finally, the combined minimum cost of distortion and bit usage is used to evaluate which combination to accept.

The labelling is for block and sub-block differs as macroblock (MB) and sub-MB in H.264/AVC and largest coded unit (LCU) and sub-CU or CU in HEVC. The crucial

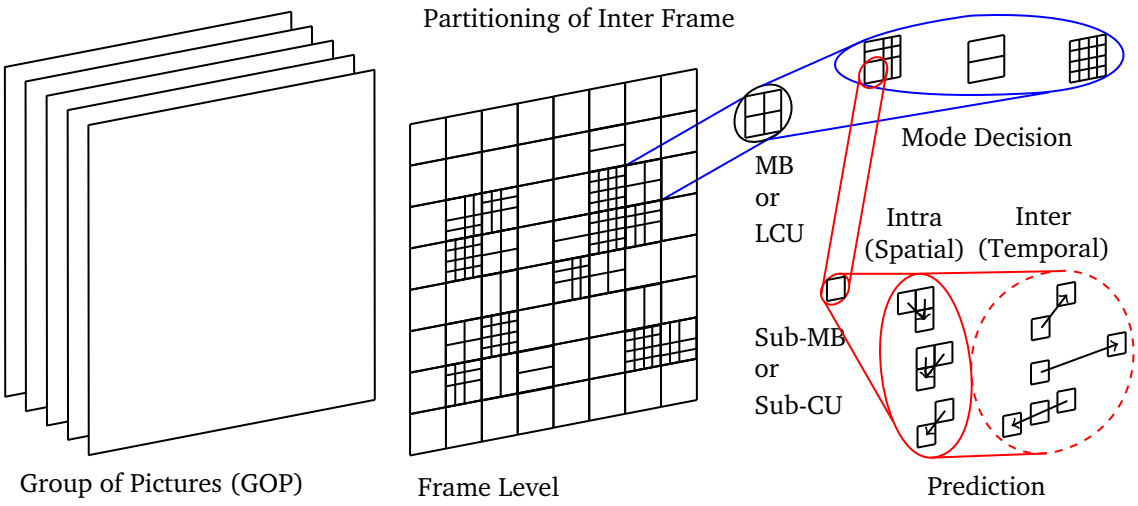


Figure 2.2 Video coding detailed view : Group of Pictures (GOP), where a series of frames will be in a collection, then at the individual frame level, finally macroblock (mode decision) and sub-macroblock (prediction) levels are shown

difference between the intra and inter block types is the size of residue and their application. Intra forms a reference from which other blocks can add or subtract changes from and provide the highest picture quality least timing, however, this has the largest bit usage. Inter come in two forms, semi-reference (predictive frame, p-frame) and no-reference (bi-predictive, b-frame), under b-frame the encoding timing increases and bit-usage decreases as less reference data is stored. This means that p-frames are for responsive critical applications and b-frames are for storage critical applications. Overall, as the number of different sub-block sizes and signalling choices increase, the volume of candidates grows, which places a greater burden on the use of distortion assessment.

2.2.2 Front-end workflow of a hybrid block-based encoder

The front-end hybrid block-based encoder workflow can be described as where each frame from a video source is divided into uniform blocks before evaluating at the (sub-) block level. Then, depending upon the configuration intra or inter prediction coding will take place and the respective distortion assessments take place as presented in Figure 2.3. For intra, different modes will be compared, for inter, motion vectors will be considered, to find the best candidate, then this process should be repeated for the next sub-block. The range of prediction candidates

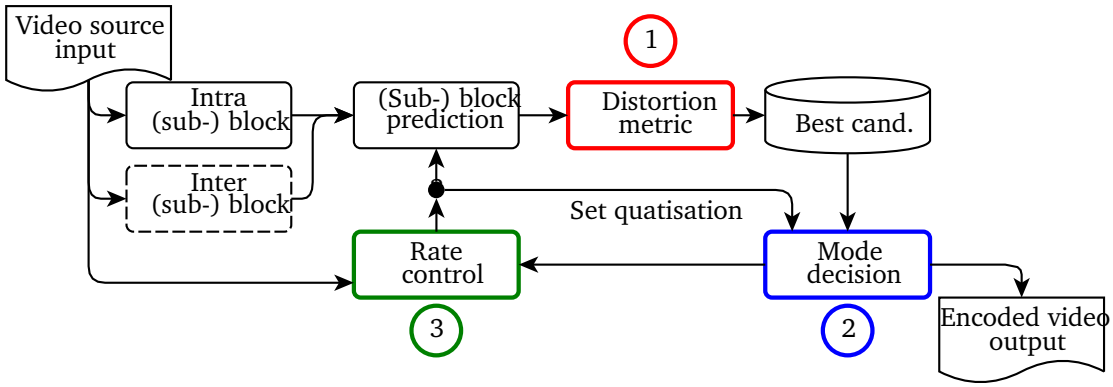


Figure 2.3 Simplified front-end video encoding workflow for a block based encoder, where numbers 1-3 indicate where a distortion metric is used

means that initial distortion metric at stage '1' is the most simplistic and extensively used. This is because the best candidates for the respective block sizes must be found. At stage '2' the best candidates for each sub-block are placed in a variety of combinations to find the best match for the given cost. During these two stages of '1' and '2' there is a search for the minimum combined cost, of bit-rate against distortion present in the (sub-) block, known as rate-distortion optimisation (RDO). Meanwhile, as distortion and bit usage can be seen as opposing forces, it must be regulated with the use of quantisation which is set by rate control, at stage '3'.

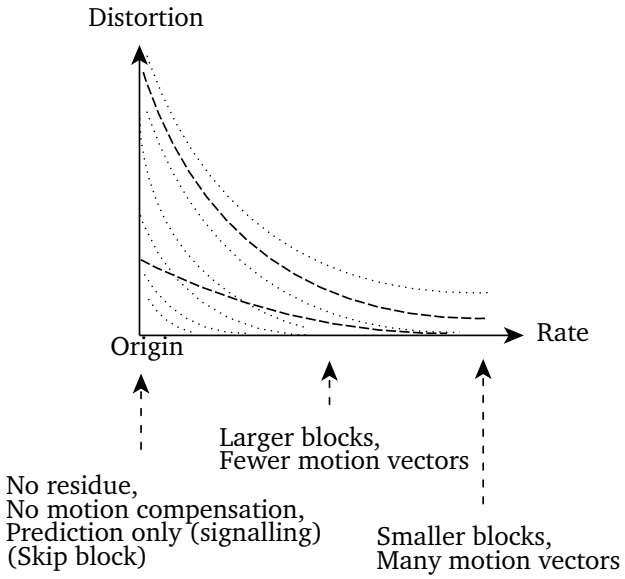


Figure 2.4 Rate-Distortion curves

2.2.3 Rate-distortion optimisation (RDO)

Balancing rate and distortion by mode decision is described as RDO and illustrated as a principle R-D curve in Figure 2.4. The RDO is where different combinations of block sizes are considered against ones which send no residual pixel difference information. RDO is most significant in inter blocks, since signalling is sensitive to the quantisation applied on the associated motion vectors, which in turn can affect block matching accuracy and increase distortion. Each mode decision option, whether a series of small blocks, a single large block or a copy from the adjacent co-located frame (without additional residual information), is considered against the R-D curve. These options can present varying degrees of differences, meaning the R-D curve changes according to the video content and signalling choices. For a highly textured video content, this may risk high levels of distortion, especially if there is movement. This effect is magnified when larger blocks with strong levels of quantisations is applied to meet a bandwidth requirement. In comparison, the R-D curve, for a homogeneous region is likely to incur lower rate of increase in distortion as larger blocks are used or even no residual information is encoded. The RDO process that produces the R-D curve can be represented as in the form of Equation (2.1)

$$J_{min} = D + \lambda \cdot R \tag{2.1}$$

where J_{min} is the total energy, D is the distortion for a given assessment sub-block, R is the bit-rate and λ is the Lagrange multiplier. λ recognises that the optimal solution for two opposing variables is measured by the one closest to the origin (Everett III, 1963). The tangent along the R-D curve represents the quantisation required for the desired bit-rate (Ortega and Ramchandran, 1998). This means that as λ strives to meet a target bit-rate this can affect the distortion passed within the encoding.

2.2.4 Lagrange multiplier and quantisation parameter (QP)

While RDO and rate control can adapt the value of λ, the encoder translates this value of λ to a Quantisation Parameter (QP) setting. Thus, an integer representation can be saved and encoded in the encoded bitstream. QP can be calculated as shown

in Equation (2.2) based upon experimentation results which shows that it can be represented with a linear equation (B. Li, D. Zhang et al., 2012).

$$QP = 4.2005 \cdot ln \cdot \lambda + 13.7122 \tag{2.2}$$

Equally, from this same investigation, while QP is a convenient way to encode the level of quantisation in the bitstream, it is not suitable to regulate rate-control at the encoder. This is because greater precision can be achieved by adjusting λ (B. Li, H. Li et al., 2012).

2.2.5 Distortion assessment

Distortion assessment occurs both at prediction and mode decision stages. During prediction stage, either SAD, or the more complex SATD is applied, while in mode decision the SSE is used. At the prediction stage, the complexity must be low as the volume of candidates are large. In comparison, at mode decision a more complex distortion is possible as fewer assessments take place. A compromise in complexity for assessment occurs with sub-integer prediction. Sub-integer prediction is an acceptance of the reality of video encoding, where the content does not align with the sub-block prediction. For that reason, values must be interpolated to obtain a reasonable estimation. An effective means of distortion assessment is the use of the Hadamard function based distortion assessment SATD. SATD is far more complex than SAD as the SATD ignores symmetrical pairs to obtain a score while being compatible with SAD range of scores. The 1D Hadamard transform is illustrated as a Tukey butterfly diagram in Figure 2.5. To perform 2D Hadamard transform, the 1D Hadamard transform of the x-axis undergoes the same process, on the y-axis.

2.2.6 Rate-control

Video coding involves focusing on maintaining picture quality or regulate bandwidth. To maintain the picture quality, selected candidates under the R-D curve would incur less distortion, yet occupy more bits, this is typically occurring in RDO. While to regulate bandwidth, the encoder seeks to limit the bits used by allowing distortion to be encoded. This is known as rate-control, where quantisation is adjusted by way of λ to meet the desired bit-rate target. Adjusting λ is more complex

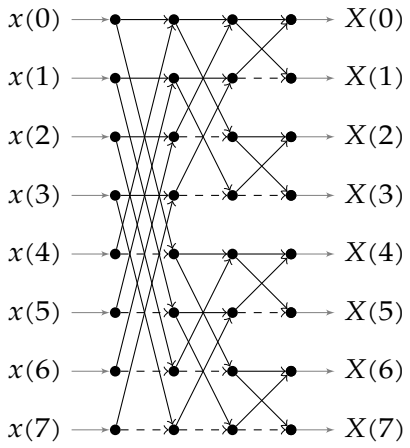


Figure 2.5 Signal flow (butterfly) diagram of 8-bit Hadamard transform, solid lines are additions, dashed lines are subtractions. Process is repeated on y-axis to perform 2D Hadamard transform.

than RDO and involves assessing activity and monitoring encoding statistics. As such, rate-control is dynamic, based upon the incoming content and the current bit usage to ensure that bandwidth restriction is adhered. Figure 2.6 illustrates the state diagram representing the rate-control process. The assessment of activity is intensive and is processed less frequently than prediction or mode decision, instead once every GOP, frame or block. However, after the initial frame, subsequent activity assessments are used in the model as weighted values. This is because the statistics of encoded blocks force the model to re-calculate and adjust λ in order to meet the bit-rate target. Overall, rate-control ensures that bits are allocated for the given video content, however, this is based upon the frame activity.

2.3 HEVC video coding standard

Among the hybrid block-based video coding standards, HEVC is the most recent and will be discussed within this section. It builds upon the previous generation

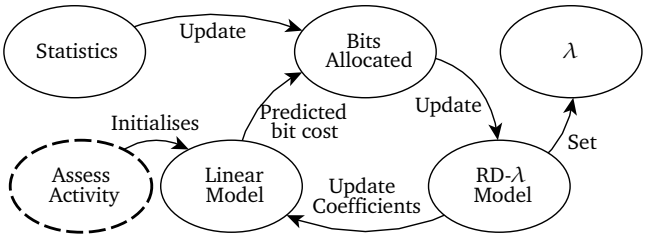


Figure 2.6 State diagram of rate control.

of H.264/AVC extending its application to low powered and/or portable devices. While the hybrid block-based principles of prediction, mode decision, RDO and rate-control remain, it has greater signalling choices which increases encoding complexity. These additional mechanisms in HEVC include greater flexibility in signalling and a distinction between signalling and quantised differences transform structures. Collectively, these all contribute towards HEVC success of bit usage reduction over H.264/AVC and will be presented in this section.

2.3.1 HEVC overview in terms of the video coding layer (VCL)

Each generation of the video coding standard has evolved to provide greater interoperability. H.264/AVC showed this with support for streaming video across the Internet and HEVC was designed to support portable devices. The underlying video coding layer (VCL) applies the same hybrid-block based compression since MPEG2, yet the VCL has developed with new terminology to meet current generation challenges (M. Pourazad et al., 2012; Sullivan et al., 2012). In all, the HEVC encoding process is largely similar to other hybrid-block encoders like H.264/AVC. This is shown in Figure 2.8, whereas in H.264/AVC, blocks are processed, involving search for intra or inter candidates from which coefficients are quantised and transformed.

The HEVC standard is designed to utilise the increased processing capabilities of processors in order to achieve a lower bit usage (Bossen et al., 2012). It achieves this lower bit usage with a series of techniques, notably by extending the hybrid-block-based codec H.264/AVC with larger block sizes and using a quad-tree structure to manage the partitioning of blocks, (Sullivan et al., 2012). In H.264/AVC the macroblock would be 16x16 pixels, in HEVC, the LCU can be, 16x16, 32x32 or 64x64. Within each LCU, the CU may be split to in order for a suitable prediction unit (PU). For intra, split is symmetrical only, while for inter, the PU can also be asymmetrically split. Regardless, prediction seeks for the best match to minimise distortion through quantisation and encourage greater use of signalling.

Another new concept in HEVC is the motion vector coding strategy, where the choice of candidates from different positions are considered, it was adopted into the standard as it reduced bit usage. This is based upon work that extends the statistical

redundancy concept from frame to block level (D. W. Dong and Atick, 1995; J.-L. Lin et al., 2013). The experiment was based upon eight reference positions (six intra and two temporal) producing mean square error (MSE) surface maps for differences based from each reference position, and the findings encouraging use of inter candidate with even distribution. As this experiment was based upon MSE, there is potential that a IQA could alter the choice of MV or use skip blocks. HEVC does have other significant changes to encourage higher compression efficiency via representation through signalling, this includes the quad-tree block structure and extending prediction to support asymmetric motion partitioning (AMP). AMP increases the choice of signalling options, however, this places greater responsibility on the assessment of these options.

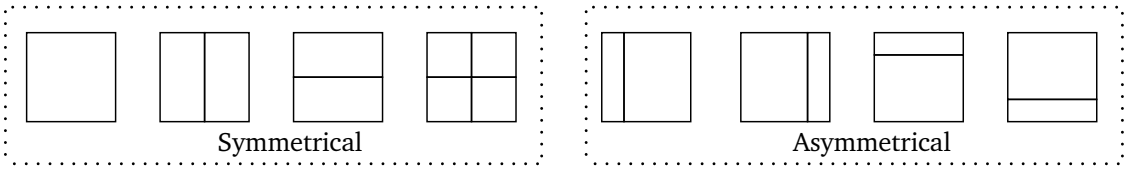


Figure 2.7 Symmetrical and asymmetrical motion prediction (AMP) modes for inter coded blocks.

2.3.2 Asymmetric motion prediction (AMP)

The quad-tree structure within the VCL continues with prediction, as PUs. For intra prediction, the prediction blocks (PBs) are the same size as the coding blocks. In inter prediction the choice of block sizes increase with the use of AMP, with AMP extending the traditional ratio choices of 1:1, 1:2 and 2:1 to more uneven splitting of block sizes with ratios of 1:4, 3:4, 4:1, 4:3 as shown in Figure 2.7. Figure 2.7 shows the existing symmetrical (represented as $M\times M$, $M\times M/2$, $M/2\times M$ and $M/2\times M2$ where M is the block width) and AMP (represented as $M/4 \times M$ for left and right and $M \times M/4$ for up and down). This allows for greater signalling choices for improved block matching to minimise the residual pixels during prediction stages.

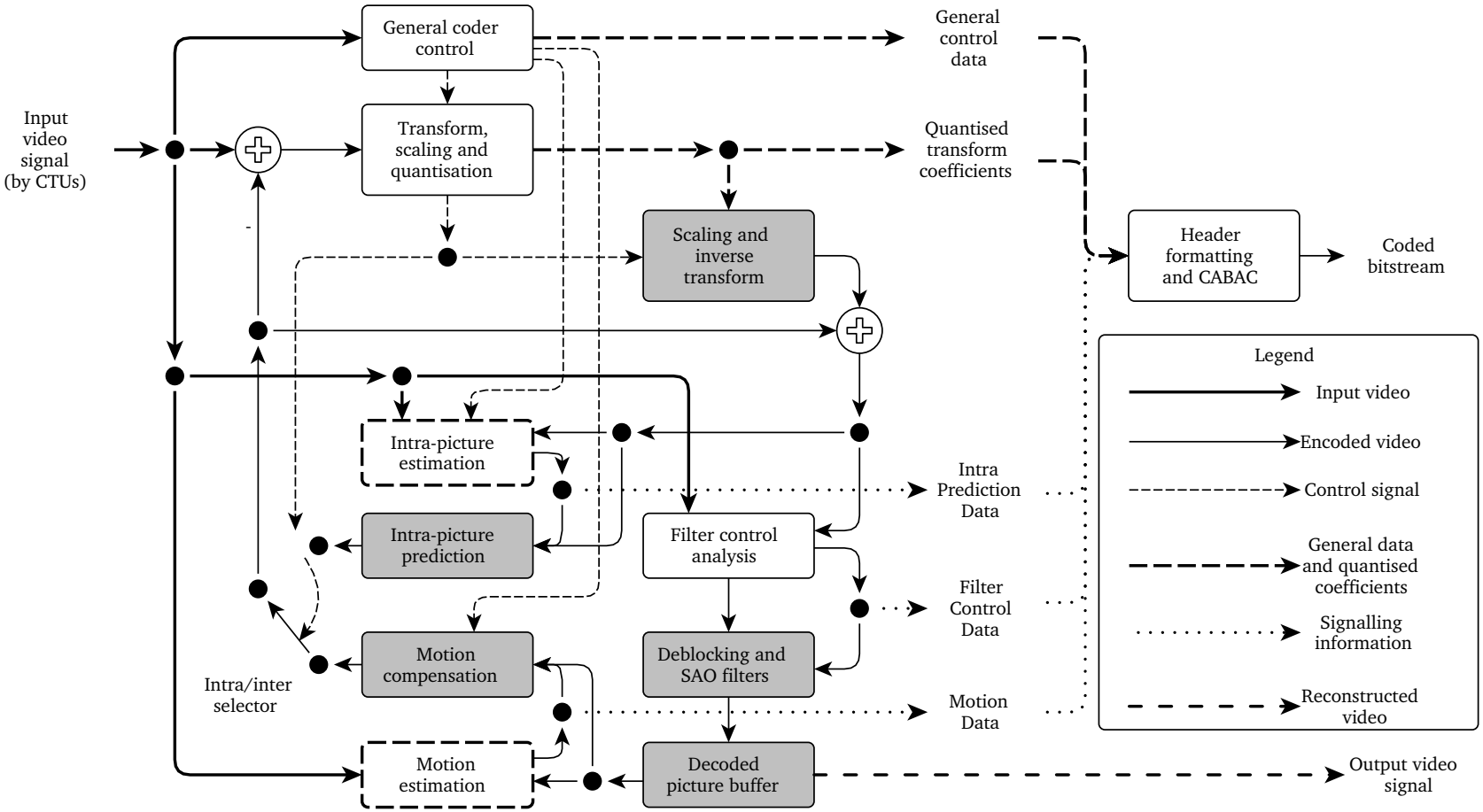


Figure 2.8 Overview of the High Efficiency Video Coding (HEVC) Standard Sulliavan et al. DOI: 10.1109/TC-SVT.2012.2221191 Typical HEVC video encoder (with decoder elements shaded in grey), redrawn with paths types identified

2.3.3 Quad-tree structure

The HEVC standard continues the use of hybrid-block based approach of operating at the sub-block level, assigning candidates based upon their minimum energy (J_{min}). However, HEVC employs a structure to manage this division of a block into its respective sub-blocks known as the quad-tree structure. Quad-tree structure is where a block is divided into four equally sized sub-blocks, and a sub-block may be sub-divided once again. This provides a symmetrical division of the block as each new sub-block level results in corresponding stub to identify where branching stops. The quad-tree structure is adopted as coded tree unit (CTU) and coding tree blocks (CTBs), with leaves of the quad-tree structure representing the partitioning known as coding unit (CU) and coding block (CB) respectively. CU(s) represent sub-blocks divisions of the LCU, which could be a mulitple sub-blocks or a single entry equal to the LCU size. CB(s) reflect that CU(s) consist of luma and chroma within the CB(s). For HEVC under the main profile, this means that for every $L \times L$ size luma CB, there will be two $L/2 \times L/2$ size chroma parts. Also, as CUs are divided, the division of the sub-block continues with the respective coding blocks, except when luma $<$ 8x8, to avoid memory bandwidth issues.

2.3.4 Transform blocks and units

The quad-tree structure is also applied to transform blocks (TB), however, unlike CTB structure based upon PBs, TBs are split, with square only sub-blocks. This further encourages TB splitting into (sub-) blocks quadrants than CTB, which means a non-square CU during AMP will be represented by multiple TBs. For every division another set of luma and chroma TBs is associated. In CTB, the minimum sized TB is 4x4, at which point luma is the same size as the chroma parts. An

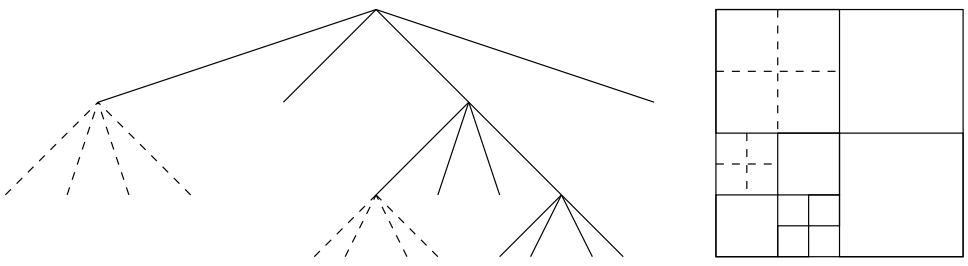


Figure 2.9 Quad-tree structure and sub-division of coded tree block for coded (CB) and transform (TB) blocks (solid lines CBs, dotted TBs)

example of this can be shown in Figure 2.9, where the z-order quad-tree structure reflects the sub-division partitioning of the block. Applying a fixed square block size for TB is due to efficient transform methods, this coding efficiency also extends to non-square PBs. Overall, quad-tree structure during TB provide greater symmetry than PB, which allows for a more compact solution.

2.3.5 Inter prediction motion merge mode

The hybrid-block based encoder is able to provide significant reductions in bit-rate by applying the principle of identifying statistical redundancy. Typically, this would occur mostly during inter coding, where similar changes are grouped or a movement from a neighbouring/previous co-located block allocated. Under H.264/AVC inter prediction would be known as direct mode and skip mode; with direct mode inferring the motion vector (MV) from the collocated blocks, and for skip mode no MV or quantised residual is encoded. In HEVC, a more refined direct mode is introduced, namely merge mode, where a fixed number of merge candidates are generated and referred to via an index. The availability of these merge candidates offer more precise and local substitutes, than the inferred approach of direct mode. These merge candidates are MVs of neighbouring PUs outside of the current CU with similar MVs to the current PU. In terms of skip mode, this is a special merge mode case, where no residual information is transmitted, only the index representing candidates which are merged. Merge mode like other advances in HEVC bring changes to offer greater precision to lower bit usage using low complexity design. The choice and selection of candidates by distortion assessment, including during merge mode, can therefore influence the distribution of bits within the block and frame.

2.3.6 Low complexity

HEVC was initially designed as two separate code bases, one aimed at broadcast and the other at real-time communications (Wiegand et al., 2010). This led to the initial HEVC test model (HM 1.0) supporting separate high efficiency and low complexity configurations (Sullivan et al., 2012). Over subsequent revisions of HEVC, these two design goals have been unified by assessing coding benefit against computational complexity. Furthermore, HEVC as a video coding standard

offers parallel computing support by splitting a frame into tiles or staggering the processing of CTUs across the available processors. Applying low complexity techniques in parallel is suited for growing number of low powered devices capable of video coding. This is important as it means that video coding can be used for the applications of broadcast and video communications in new environments.

2.4 The Human Visual System and modelling the initial V1 stage

While the previous sections discussed the encoder these following sections will describe the Human Visual System (HVS), the means of perceptual evaluation and coding techniques. The HVS is highly complex, with different aspects being investigated on multiple research fields, including investigating sensitivity to changes in lighting, detection of edges, recognition of shape, distinguishing textures and tracking movement. Overall, the HVS is classified into several stages to represent higher levels of processing, however, as the understanding increases these stages are further subdivided. For the purpose of this section the focus is on the initial HVS stage, the retina and when the primary visual cortex of the brain starts to handle visual information.

2.4.1 The retina

The HVS initiates from the retina, which lies at back of the eye, collecting visual sensory information. The retina is considered as an extension of the brain as tissue from the retina directly connects the capturing of light through to the primary visual cortex, located on the lower back part of the brain (Kolb, 2003; Schmolesky, 2016). In the retina, light must go through the 6 major layers before reaching the light sensitive rods and cones used to detect light and colour respectively. The rods aid in low light vision, allowing slow adaptation to dark environments, while rods are colour sensitive either red, green or blue, with the highest concentration at the fovea, where the image is focused. The colour sensitive rods of red, green and blue refer to short, medium and long wavelength respectively. This means that the HVS is sensitive to lighting changes and less so to colour unless at it is acuity of the image. In terms of video coding, limitations in coding has typically meant allocating greater bit depth to luma and less so to chroma parts. However, with

HEVC, support for HDR and wide colour gamut means manufacturers, content providers and broadcasters are providing increase bit-depth to provide a more emersive experience (Y. Dong, M. T. Pourazad and Nasiopoulos, 2016).

2.4.2 V1 stage

The retina converts light to bio-electrical information across the lateral geniculate nucleus (LGN), enabling visual information to the V1 stage of the brain, known as 'the striate cortex', via optical nerve fibres (Carnec, Callet and Barba, 2008; Schmolesky, 2016). The primary visual cortex is made of multiple visual stages (V1 to V5 and media-temporal (MT)), that feed information between these stages. Each stage can have several layers within it. For example, in V1 stage there are 6 main layers, yet as greater understanding is gathered more sub-layers are identified. In terms of the initial V1 stage, nerve fibres can transmit information from the retina to different layers in the V1 stage. In all, the first layer in V1 is classed as a network from which other more dense layers (2-5b) have pyramid like links to the other HVS stages of V2 to V5 and MT. Generally, the V1 stage provides a pivotal role in processing of visual information to allow higher layers to discriminate, recognise and track objects.

2.4.3 Modelling the V1 stage

As highlighted above, the V1 stage is crucial, as it perceptually filters visual information for the later HVS stages where a topographical sense map is produced. Consequently, being able to model the V1's behaviour can provide a means to present visual information suitable for the HVS, while experiments for human perception have occurred, it this remains difficult fully explain the HVS (Chandler, 2013). Investigations to model the HVS, in particular the V1 stage have shown that human perception to change is non-linear, dependent on frequency (Cowdrick, 1917; Weber, 1864). This understanding has been applied to image processing, as means to utilise bandwidth due to Shannon's rate-distortion theory, thus propelling the need to seek perceptual means to measure distortion non-uniformly in order to achieve lower bit-rate (Mannos and Sakrison, 1974). This has resulted in the non-linear frequency based model called Contrast Sensitivity Function (CSF), which can identify regions that are perceptually significant. Potentially, for video encoding

being able to apply CSF during quantisation can allow bandwidth reduction whilst maintaining perceptual quality. However, as the V1 stage is still being understood, models such as CSF are only approximations and yet are highly complex to implement.

2.4.4 Understanding of the V1 stage

Experiments for measuring the V1 stage have led to the CSF model, where the findings revealed that HVS has a non-linear threshold sensitive to frequency (Peli, 2001). However, the respective threshold is relative to local conditions, allowing for greater dynamic range to be supported by the HVS. This means that below these thresholds the HVS is not sensitive to level change in contrast, yet upon reaching these thresholds, the non-linear HVS sensitivity response occurs. Overall, this description is simplistic, and as it is claimed that up to 85% of the V1 stage has yet yo be defined, which highlights the HVS complexity (Chandler, 2013).

2.5 Evaluation of models for perceptual video coding (PVC)

The HVS is complex, and its understanding is limited, the perceptual models that have been produced are based upon this limited understanding. These models can be applied to both image and in video coding, and has led to the need for both objective and subjective testing. In image coding, database of images with different levels and types of distortion was introduced when SSIM was launched (Z. Wang et al., 2004). This acted as a means to perceptually evaluated non-/perceptual coding performance. There are other similar image coding databases, however, they are designed for different purposes and also regulate access to discourage training on the test material. They have in turn been used as a guide to simulate errors on various content, however, their range of video sequences (both in content and resolution) can be limiting. Similarly, in video coding, access to HD video sequences can be restricted, which does hinder development, as researchers need access test video sequence. Fortunately, an institution in multimedia development have provided new content for video coding development (University, 2016). These recent changes are welcomed, as the move towards higher resolution and larger bit-depth was being hindered and allows repeatability of experiments by others.

Ideally, video sequences should be ground-truth, however this is an exhaustive process (Chandler, 2013).

2.6 Subjective testing

Video encoding mechanisms are usually evaluated by their objective measures of PSNR or/and Structured Similiarity (SSIM), (more on SSIM in section 2.12 Perceptual assessment using SSIM). This provides sufficient means to guide how a video encoder is performing. However, as PVC is aimed at the HVS, then subjective testing, which is a measure based on the feedback of participants, is the ideal approach to rate an encoder's performance. This form of evaluation is important especially for PVC solution which takes into account the HVS when encoding. Subjective testing comes in two major methodologies, single or double stimulus. For single stimulus, the video sequence is shown once (the original or proposed encoder), and for double stimulus, the reference uncompressed video sequence is shown followed by the proposed encoder. Immediately, this means the double stimulus tests take twice as long, however, this allows for greater scrutiny (H. R. Wu, W. Lin and Ngan, 2014). Unfortunately, fatigue by the participant in such experiments means that the guidelines recommend breaks after approx 20 minutes (Pinson, Janowski and Papir, 2015; VQEG, 2004).

Another aspect which should be considered is the rating scales, which allows the mean opinion score (MOS) to be calculated. The absolute category rating (ACR) allows labels of 'bad', 'poor', 'fair', 'good', and 'excellent' to be applied which translate to numbers 1 to 5 (ITU-T, 2008). Irrespective of the stimulus method

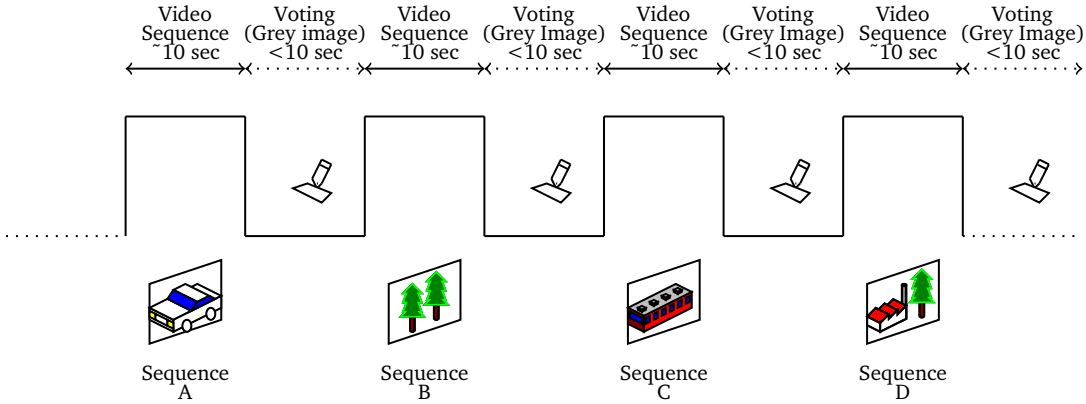


Figure 2.10 Absolute category rating (ACR) stimulus presentation method

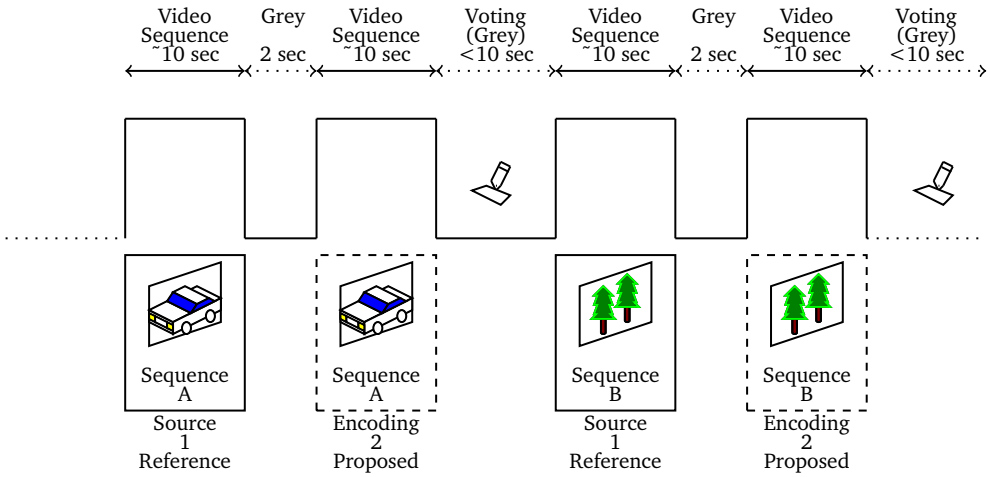


Figure 2.11 Degradation category rating (DCR) stimulus presentation method

chosen, it is recommended that video sequence of 10 seconds are used, however, it has been shown that video sequence of 5 seconds should be sufficient (Moss et al., 2015). Under ACR following each video sequence, a voting period of ≤ 10 seconds be provided, as shown in Figure 2.10. As the ACR shows video to be independently assessed, this may be unsuitable where image quality is very similar. For that, the degradation category rating (DCR) is where a source and sequence under test are presented 2 seconds apart, followed by voting as shown in Figure 2.11. To allow further discrimination when participants assess video sequences it is possible to extended the 5 point scale to 9 or 11, which would counter a criticism of 5 point scaling system (H. R. Wu, W. Lin and Ngan, 2014). Yet, since an individual is comfortable with a 5 point scale, any greater resolution will be stressful for the participant (Pinson, Janowski and Papir, 2015). There is a third configuration similar to DCR called pair comparison (PC), where pairs of compressed video sequences are used. This is where both sets processed of video sequences are shown, original encoder and proposed encoder, with the participant selecting which they prefer. In DCR, subjective differences are measured absolutely against a reference, under PC, the differences are measured relatively to each other. Finally, it is stressed that validity of the MOS score is based upon a sample size of between 15 and 40 for the evaluation to be statistically valid (ITU-T, 2008). This is stated where statically there will be little or no change, suggesting that the central limit theorem is applicable.

2.7 Video sequences for use in developing and testing PVC solutions

A key component for both evaluation and development of PVC solutions is to have access test material by way of video sequences. Such video sequences provide the basis by which others can replicate the same experiments. For standard definition (SD) resolution, this consists of raw uncompressed YUV 8 bit (4:2:0) video sequences and access to these videos are available on public websites (Xiph.org, 2016). This includes a limited HD content and thus makes it difficult to apply extensive testing. There are other databases which are non-public, and access to them is via an agreement form being signed stipulated the terms of use. Because of obstacles such as this and the need to operate at higher resolutions and bit rates, Shanghai Jiao Tong University has produced its own material and have made it publicly available (University, 2016). Equally, being able to playback raw YUV video requires specialist software, and this is possible with PYUV 'raw video sequence player' (Baruffa, 2016).

2.8 Existing perceptual techniques

The limited HVS understanding has allowed perceptual techniques aimed at image or video coding which have been evaluated. Ideally, assessment of distortion or activity in video encoding should occur using perceptual means, whereby the ranking of candidates are measured against the annoyance of its distortion than just the difference. This is acceptable, however, the means by which to undertake this challenge remains open, especially as existing understanding of the HVS is limited and existing models are highly complex. In turn, while the need for perceptual assessment remains, for pragmatic reasons, distortion continues to be assessed using STDMs, to keep the complexity envelope low (Chandler, 2013). This dilemma has not stopped the development of applying HVS models in video encoding to produce PVC solutions.

Overall, HVS based modelling or assessment can be classed under three main categories, each providing different approach to bring HVS to PVC. The three categories can be described as: sensitivity detection (low level HVS), cognitive

perception (high level HVS) and statistical based model of HVS environmnet. Low level HVS refers to the initial stages of the HVS, where filtering occurs during V1 stage. The high level HVS is where the cognitive perception is able to recognises shapes and patterns. While the statistical based model of HVS environment, states that rather modelling the HVS, models should be based upon the environment which the HVS has evolved within. Each of these have their own approach to modelling aspects of the HVS and will be discussed below.

2.9 Low level HVS models

Low level HVS refers to the early experiments which attempt to measure the response to single point sources (Cowdrick, 1917; Weber, 1864). These experiments led to producing the HVS model based upon the frequency gratings called Contrast Sensitivity Function (CSF) and those based upon luma called Just Noticeable Difference (JND) (Mannos and Sakrison, 1974; Yogeshwar and Mammone, 1990). These models depict that HVS has a non-linear sensitivity response to the threshold of perceptual detection. This means that for an equal proportion of change, the HVS will struggle to detect changes in a darker region of an image compared to a mid to brighter regions. Furthermore, the contrast formed by an object boundary or for a given texture has led to forming of edge detection. Each of these types of low level HVS attempt to filter what will get passed along the optic nerve.

2.9.1 Contrast Sensitivity Function (CSF)

Perceptual based coding harnesses the understanding of how the sensory and cognitive regions operate to identify redundancy in audio and video processing (H. R. Wu, Reibman et al., 2013). Overall, HVS models are shown to model against Weber's law (Weber, 1864). One particular model, Contrast Sensitivity Function (CSF) is based on a series of frequency-banded experiments that illustrate the non-linear threshold based response of the HVS (Kelly, 1979). More so, the CSF demonstrates that the HVS is a frequency-based filtering system subject to the relative background intensity. Unfortunatley, applying a frequency based filter to the front-end video encoder is not ideal as it is pixel based and would require transforming to the between pixel and frequency domains. Several other types of perceptual coding techniques exist and offer approaches, however, these are

liable to high levels of complexity with respect to existing non-perceptual methods (H. R. Wu, W. Lin and Ngan, 2014).

2.9.2 Just Noticeable Difference (JND)

Another perceptual coding approach is JND which allows the encoder to remain in the pixel domain as it incorporates Weber's law (Yogeshwar and Mammone, 1990). JND applies Weber's law on a basis that human perception to change is threshold based, which in this case, is relative to the pixel intensity levels. Again like CSF, JND is a filter, however, JND is in the pixel domain which is more suitable for video coding. Despite JND operating in the pixel domain, its processing requirements are high that leads to an optimised version of JND, for 8 bit luminance channel (Chou and Y.-C. Li, 1995). Regardless, CSF and JND are HVS models, indicating the sensitivity to distortion on a given region before it may be perceptually detected. Neither CSF nor JND quantify as an IQA as they operate with only the original image and are not designed to measure distortion of the differences between the original and reconstructed images (H. R. Wu, Reibman et al., 2013). This means that while CSF and JND can provide indications of sensitivity, they are unable to measure perceptually significance of the distortion.

2.9.3 Edge detectors

A different approach to low level HVS is edge detection, it produces a binary result, applied to the original image. While this does not make edge detection an IQA, nor a perceptual model like CSF and JND, edge detection can indicate where boundaries for objects or textures exist. Edge detection examines the connected nature of adjacent pixels in an array for whether the rate of change exceeds a threshold, then an edge is said to be detected. This is possible by an edge detection operator applying a square block size around pixel test point to detect changes against a threshold. This means that if the threshold is too low, it is liable to suffer from false-positives and conversely, if the threshold is too high the edge detection fails to detect (Robin E. N. Horne, 1998). In addition, edge detections have high complexity making them cumbersome, especially as they require several test points for a single edge detection, which can be shown with the Sobel edge detection (Kay and Lemay, 1986). This means that while edge detection has a unique means of

evaluating, the issues of reliability and complexity need to be addressed. Also, edge detection as a stand-alone measure is unsuitable due to it binary response and should be considered as part of series of perceptual techniques.

2.10 High level HVS models

While the low level HVS models discussed perceptual models based on the V1 stage, it is at the later stages of V2 and higher where the high level HVS models are based. Here, visual information is received following the V1 stage and objects or textures are recognised then tracked. This places additional burden on the brain to process and store vast amount of visual information. As such, high level HVS models aim to identify regions of interest (ROI) based upon attention models or by visual saliency techniques (Duncan and Sarkar, 2012). These attention models are designed around understanding how a given scene may be read by the brain and how this may translate as annoyances due to errors in transmission. There are two main forms by which this can occur; one is by synthetic completion and another is visual saliency. In synthetic completion, both the encoder and decoder are modified to store perceptually related clues (Ndjiki-Nya et al., 2012). For visual saliency, feature extraction or pattern recognition is used to identify items, which acts as more robust approach than edge detection. This makes visual saliency, computationally expensive to compute as it is more complex than existing edge detection. Equally, using synthetic completion results in a non-standard video encoder as the decoder must be modified to support the additional meta layer information which is transmitted.

2.10.1 Attention model

The HVS is a highly complex process and the current understanding is limited. However, in general, the HVS is considered as a series of stages from V1, where filtering like in CSF or JND occurs, to higher stages of V3 and V4 where recognition and tracking happens (Le Callet and Niebur, 2013). Each of these stages filter the visual information to a reduced set, relative or context-based, to allow information processing to occur within the HVS. This means that the HVS has evolved to allow feature recognition in order to survive.

In all the HVS understanding can be explained as applying a masking effect (during the initial V1 stage) to producing a feature map (V2 stage) before object recognition and tracking (at higher stages of V3 and V4 respectively) (Carnec, Callet and Barba, 2008). Incorporating these models based around the higher stages of HVS can enable adaptive region based video coding, though this comes at the expense of very high complexity.

Another way to understand the HVS is by measuring the response of participants to a level of annoyance in video coding errors. An experiment in which specific errors were applied separately and in combinations was performed (Péchard et al., 2007). This highlighted that by minimising the most critical type of annoyances, the other forms of distortion may be perceptually insensitive or tolerated. As this approach was dependent upon the scene, this can have a bearing on the design of attention models, where perceptual-related meta information for a scene is used during encoding (Le Callet and Niebur, 2013). This means that attention models enable the tracking of objects at the expense of higher complexity.

2.10.2 Synthetic completion

Another form of perceptual assessment is synthetic completion, this uses image analysis to evaluate how existing textures may be reused, which relies upon sending additional meta information to the decoder. This approach extends exploiting statistical redundant information as it introduces a new means of assessment. However, applying synthetic completion for block based video encoders of H.26x series would render them non-standard, as the decoder will need to be modified (Ndjiki-Nya et al., 2012). Synthetic completion proposes a content aware model to identify perceptually relevant textured regions suitable to be synthesised. This can consistent of combining different approaches (partial differential equation and non-parametric), making the solution able to distinguish edges from texture. When implemented, synthetic completion has shown bit usage savings of up to 40% on a HD source. Unfortunately, little was given in computation complexity. Similarly, the authors emphasised that ideally a perceptual distortion metric should be used to dynamically detect ROI.

2.10.3 Visual saliency

Underlying the need to apply ROI is the use of visual saliency, which can identify patterns and track objects, by connecting and recognising the various edges or boundaries detected. Incorporating visual saliency into a perceptual distortion can be convenient in terms of a concept for video coding, however, computational loads are high (Duncan and Sarkar, 2012). This is irrespective of which of the three types of visual saliency that is used, memory-less (bottom-up), prior knowledge (top-down) or limited knowledge (integrated). Each of these visual saliency types refers to filtering information, while the last two are also able to recognise features. Regardless, of how each visual saliency algorithm differs they are generally computationally demanding, where even bottom-up solutions are recommended to be optimised (Duncan and Sarkar, 2012). Overall, this means visual saliency is not a preferred option for the use in PVC.

2.11 Statistical based model of the HVS environment

Another means to HVS modelling is a counter approach, where instead the natural environment which the HVS has evolved around is modelled based upon the statistical information present (A. C. Bovik, 2013). A key difference of statistical based HVS models over low level HVS models is that distortion can be measured perceptually, providing a means to compare performance with STDMs. This is because statistical based HVS models are able to consider the relationship of original and the reconstructed with respect to a HVS model. This distinguishes the statistical HVS as one where they can be applied as a form of: image quality assessment (IQA), texture similarity or mean opinion score (MOS). In the real world of everyday life, the HVS applies assessment without a reference image, yet for image and video encoding systems a reference image is used. From these encoding systems perspective, perceptual distortion assessment is an attractive goal, since distortion measured on its perceptual merit can affect the encoders choices. Statistical based models of the HVS environment evaluate distortion perceptually, classifying it as an IQA. A popular example of this is the structural similarity (SSIM) algorithm (Z. Wang et al., 2004). Overall, this means unlike other

perceptual models which limit themselves to content alone, statistical based model of the HVS environment can be applied to image and video encoding systems.

2.11.1 Image quality assessment (IQA)

As calls are made for distortion to be measured by utilising HVS understanding, these new series of efforts are for perceptual assessments to be called image quality assessment (IQA). This means that an IQA can place emphasis on salient regions, allowing greater sensitivity to perceptual significant differences compared to the uniform cost by STDMs (Le Callet and Niebur, 2013; H. R. Wu, Reibman et al., 2013) Applying an IQA is ideal, where all distortion can be perceptually assessed, which would choose perceptually friendly candidates. While in general, IQAs cover perceptual forms of distortion assessment, IQAs can be broken down into their respective classification based on how they operate (Chikkerur et al., 2011). This includes how the IQA operates by frequency or pixel domain, and even by which methodology, such as natural scene statistics (NSS). Understanding which domain, where in the encoder it operates and how it was devised provides an indication of its application and limitations. Amoung these IQA solutions, SSIM, a pixel based IQA based upon NSS has been shown to be the least complex and best performing compared to other IQAs (W. Lin and Kuo, 2011). Unfortunately in terms of video coding, IQAs like SSIM do not provide an elegant solution, yet SSIM will be revisited later in section 2.12 Perceptual assessment using SSIM. This is because, while IQAs provide scores to assess distortion, they themselves operate using methods which make their complexity undesirable and also incompatible with existing STDMs. However, this does not stop the development of IQAs with higher complexity.

2.11.2 Texture similarity

The development of IQAs continue to evolve, including increasing the perceptual accuracy by applying more complex forms of perceptual assessments. Texture similarity is one example, where the NSS approach is extended beyond the limit of 2^{nd} order of statistical moments (T. N. Pappas et al., 2013). This is where pixels are seen as connected regions and attempts are made to perceptually synthesise them. Underlying this approach is the Julesz conjecture, where an image texture is represented by its statistical moments, this can indicate the use of higher than

V1 stage of the HVS (Julesz, 1981). Texture similarity draws inspiration from NSS and produces an algorithm that can synthesise textures with varying degrees of success. This means that in video coding the signalling information will contain greater perceptual significance. Potentially, this can improve compression levels while maintaining structural integrity, though with higher statistical moments comes greater complexity.

2.11.3 Mean opinion score (MOS)

The use of people to evaluate video is a means by which subjective evaluation can occur, however, when their input is used to create a model, this presents a time-consuming yet interesting experiment. This approach was applied to produce a PVC solution which would use the MOS of participants based upon results gathered from different videos (Bhat, Richardson and Kannangara, 2010). Consequently, this produced a solution aimed at HVS by creating a model using different HVS based tests, however, limitations of the experimental set-up and implemented solution make it unattractive. The results were encouraging, however, they were of common interchange format (CIF) (352 by 288 pixels) resolution, a quarter of standard definition (SD), which limits its appeal. Overall, the approach was unique, relying upon participants to contribute towards a model, yet the design is highly complex.

2.12 Perceptual assessment using SSIM

This section is specifically aimed at the SSIM algorithm since it is widely accepted as a perceptual alternative to PSNR and used as an IQA in PVC solutions. This makes SSIM attractive to use since it provides a single score which is ideal. A single equation to evaluate distortion perceptually potentially overcome the issues related to mixing perceptual models and techniques. This would mean that the scaling of issues would have been addressed at algorithm design stage rather than during implementation stage. Furthermore, while multiple perceptual models and/or HVS understanding can assist to produce a more robust PVC solution, it means the additive complexity of applying multiple perceptual techniques. This problem of greater complexity when multiple perceptual techniques are combined, means that an integrated solution is required. Potentially, an integrated solutions should offer perceptual robustness within a complexity competitive envelope

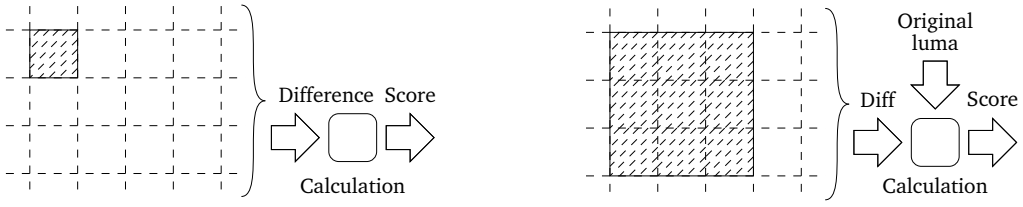
compared to STDMs. This issue was highlighted with structural similarity (SSIM) where perceptual aspect of luminance, contrast and structural comparison were represented in the form of the SSIM equation (Z. Wang et al., 2004). The SSIM equation which is presented as an alternative to PSNR to measure distortion has difficulty in being adapted for replacing STDMs in video encoding. SSIM presents a unification of three perceptual measures, however, the scaling of SSIM score to STDM is difficult. This can be define as an incompatibility of SSIM against STDMs within the video coding environment, which requires an understanding how SSIM and the respective STDM (Richter, 2011).

2.12.1 SSIM equation

Since its inception, SSIM has been significantly investigated and adapted in different image and video coding scheme as an IQA. It differs from existing distortion assessments since it factors the original luma values and calculates local statistics via a sliding window. This is shown in Figure 2.12, where the existing distortion assessment is accumulated from the individual pixel differences. In comparison, in an IQA like SSIM, a sliding window is used, covering the neighbouring pixels to calculate relative differences against the original to produce the SSIM score.

Pre-SSIM

The motivation for SSIM is centred around the NSS which identified the need to assess distortion perceptually, however, it is unclear how exactly the SSIM equation was eventually formed. The SSIM algorithm resembles the findings of experiments undertaken earlier on the perceptual effects of luma changes, in particular, the choice to normalise as an index range between ± 1, as shown by the algorithm



(a) Existing distortion (b) Perceptual assessment (SSIM)

Figure 2.12 Existing distortion vs sliding window based IQA (SSIM)

labelled 'structural similarities' in Equation (2.3) (Caelli and Moraglia, 1986; Chandler, 2013; Field, 1987).

$$\frac{\sigma_{xy} + C_3}{\sigma_x \cdot \sigma_y + C_3} \approx \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{Cov(I_1, I_2)}{\sigma_x \cdot \sigma_y} \tag{2.3}$$

where σ_x, σ_y) is standard deviation for x and y, $\sigma_{x,y}$ is covariance and C_3 is a constant equal to $(0.03 \cdot L)^2/2$, where L is dynamic range of bit depth. This equation is reflective of the SSIM's structural function. Equation (2.3) produced an improved perceptual performance, though the authors added caution, due to the increased false alarms rates (Caelli and Moraglia, 1986).

Calculating local statistics

SSIM is based upon the fact that HVS has evolved around the natural world of NSS. The calculation of statistics of distortion for IQA indicates towards the 'Julesz conjecture'. The 'Julesz conjecture', states that an image texture can represented by its statistical moments (Julesz, 1981; T. N. Pappas et al., 2013). Unfortunately, capturing image texture with higher orders of statistical moments become computationally expensive, suggesting greater diminishing returns for greater than 2nd order statistical moments. SSIM encompasses statistical moments in its calculations of the three perceptual measures; luminance (l), contrast (c) and structure (s), as presented in Equations (2.4) to (2.6) respectively. These are combined in the form of Equation (2.7), eventually resulting in Equation (2.8)

$$l(x,y) = \frac{2\bar{x}\bar{y} + C_1}{\bar{x}^2 + \bar{y}^2 + C_1} \tag{2.4}$$

$$c(x,y) = \frac{2\sigma_x \cdot \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{2.5}$$

$$s(x,y) = \frac{\sigma_{x,y} + C_3}{\sigma_x^2 \cdot \sigma_y^2 + C_3} \tag{2.6}$$

$$SSIM(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma} \tag{2.7}$$

$$SSIM(x,y) = \frac{(2\bar{x}\bar{y} + C_1) \cdot (2\sigma_{x,y} + C_2)}{(\bar{x}^2 + \bar{y}^2 + C_1) \cdot (\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2.8}$$

where the image statistic moments (in terms of mean ($\bar{x}$, $\bar{y}$), standard deviation (σ_x, σ_y), variance (σ_x^2, σ_y^2) and covariance ($\sigma_{x,y}$)) are used to calculate SSIM. While C_1, C_2 and C_3 are constants added to ensure stability. $C_3 = C_2/2$, leaving C_1 and C_2 to be represented in the form of $C_n = (K_n L)^2$ where L is dynamic range of bit depth, for 8 bit, 255, K_1 is typically 0.01 and K_2 is typically 0.03. This understanding can

allow pre-calculated values to be called, rather than for the values to be calculated each time.

2.12.2 SSIM in relation to HVS models and STDMs

SSIM was designed to be a reliable perceptual measure for distortion, where distortion can take one of two paths; perceptual or non-perceptual as part of a 'unit-circle' (Z. Wang et al., 2004). The SSIM equation defines the purpose of an IQA to treat perceptually undetectable differently to those which are perceptually annoying. SSIM does this by taking the statistical properties of the differences to calculate the relative perceptual impairment. This approach means that changes are evaluated against neighbouring local pixels than individual pixel differences. The advantage of SSIM is that among IQAs, SSIM offers the least complex of the major alternatives and is suited for block based encoders (W. Lin and Kuo, 2011). While SSIM is commonly used in PVC solutions, it is more complex than existing distance measures, yet less so than perceptual models, as shown in Figure 2.13. The complexity associated with SSIM may be less than its peers, however, SSIM is a PVC solution which requires additional complexity to ensure compatible scores with existing STDMs.

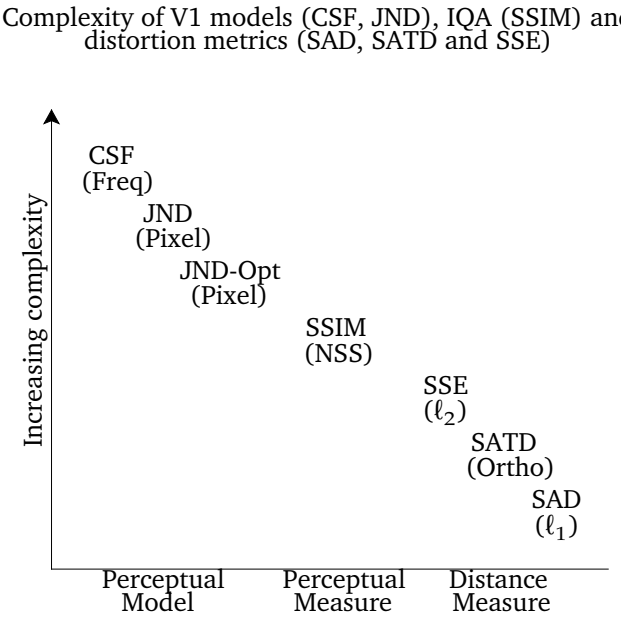


Figure 2.13 Complexity: SSIM vs Perceptual Models (CSF and JND) and existing Distortion Metrics (SAD, SATD and SSE).

2.12.3 Suitability of SSIM in video coding

SSIM operates by providing a perceptual measure for distortion between a pair of original and reconstructed images, known as a full reference IQA (Z. Wang et al., 2004). The HVS models of CSF and JND are filters, which approximate the initial V1 stage of the HVS based upon an original image. This means they indicate areas of sensitivity, however, they do not assess perceptual significance of the distortion. Also, the CSF or JND models operate at the frame level, while SSIM operates using a sliding window, which is more suited for the hybrid block based encoder of H.264/AVC or HEVC. However, due to the complexity associated in scaling SSIM to STDMs, SSIM is applied in PVC solutions where quantisation is perceptually adapted, resulting in a perceptually rescaled quantisation (λ_p) based solution. This means that a model is applied to candidates retrospectively, rather than candidates being assessed individually, which is discussed in detail in Section 3.4 Existing PVC solutions.

2.12.4 SSIM vs. PSNR

Traditionally, PSNR has been used in image and video assessment as an objective measure, where the numerical result is considered representative of the losses. Over time, PSNR has had questions of its effectiveness and SSIM has gained attention as an objective perceptual alternative to PSNR based upon studies including MOS (Z. Wang et al., 2004). In a more detailed study, SSIM was mathematically analysed against PSNR, showing that the relationship was linear along a distinct range (SSIM of 0.2 to 0.8) (Horé and Ziou, 2013). Their approach was straightforward; applying four different types of distortions each with increasing noise parameters, finding that SSIM was consistent across all forms of distortion, whereas PSNR was prone to Gaussian noise. This illustrates that SSIM has an ability to ignore spatially uncorrelated pixel values due to the use of NSS, while for PSNR, with large differences on small number of pixels can affect the score. However, this does mean that SSIM does ignore changes which might be averaged out due to the use of windowing.

SSIM windowing

The use of a sliding window in SSIM is related to calculating the local image statistics. This process places a high computational demand and is further affected by the degree of overlap of the sliding window against the next SSIM iteration. If this overlap is the maximum, then the window will move a single pixel, leading to higher complexity, however, if there is no overlap the additional complexity is minimised. Another issue is the size of windowing which has been investigated, and shows that the preferred size is 8x8 (Brooks, X. Zhao and T. Pappas, 2008; Z. Wang et al., 2004). While smaller windowing sizes are possible, it poses the risk of a single value dramatically affecting the final SSIM score.

2.13 Integrating HVS model(s) within a PVC solution

PVC solutions need to address complexity and provide suitable PVC capability, this has led to combining multiple HVS models and later producing a single algorithm. The use of combining existing PVC models was initially centred around application specific cases, however, later this grew to become one of general use. This highlighted the inefficient design and with the advent of a single algorithm, by way of SSIM, which suggested an elegant solution may be possible. Overall, this has shown PVC has evolved to support more scenarios, however, these PVC solutions are not computationally efficient compared to existing STDMs.

Perceptual offers the potential to lower bandwidth requirements for similar picture quality, the use of perceptual assessment such as SSIM in video coding is a case of ‘managed complexity’. This has meant that when perceptual assesment is regulated when and where it is used to ensure that complexity overhead is not applied less frequently than STDMs. This management of complexity can occur at frame or block level, to manage the overall complexity (Dai et al., 2014; Y.-H. Huang, Ou, Su et al., 2010). However, applications for IoT and 6LoWPAN are propelling the need for low complexity solutions and video coding is no exception. While an STDM competitive solution does not exist, adapting existing perceptual assessment to fit an increasingly smaller complexity envelope is proving problematic. Thus, designing any new perceptual assessment without factoring complexity, risks being unsuitable for portable or low powered applications.

2.14 Application specific HVS models

The use of HVS models as part of a PVC solution was initially used to described visual importance and to regulate quantisation between these areas. This is shown in video applications for both dynamic and static backgrounds using edge detection and region of interest (ROI), respectively. The dynamic ability of edge detection means that as the content changes, the encoder can be steered accordingly. While for static videos sequences like video conferencing, ROI is applied on designated areas.

2.14.1 Using edge detection to generate an importance map

Edge-detection is a means to provide a low level HVS assessment, however, it is liable to have phantom edges when the threshold is too low, equally, no edges are detected when the threshold is too high (Robin E. N. Horne, 1998). In HVS terms, edge detection offers low processing overhead, making it attractive to use as part of a combination with other HVS models. An example of this is the Sobel operator, which aids to classify the block texture type and contribute towards the importance map for adjusting λ (Yu et al., 2005). This allowed an importance map to be formed, which enabled λ to be adjusted during the mode decision as part of rate-control. The PVC solution is aimed at identifying adjacent frames with high motion and then using the importance map to affect quantisation at the macroblock level. The design limitation of this solution is the Sobel operator, which covers a region of 3x3 (Kay and Lemay, 1986). This means that edge detector at the sub-block level would prevent the sub-block perimeter being tested. As the sub-block perimeter may have discontinuities, the potential to influence the perceptual performance is reduced with the Sobel operator. A smaller edge detection is currently not available which would enable more test points to be covered during smaller sub-blocks. In turn, a smaller edge detection would enable a higher resolution importance map to be generated.

2.14.2 Region of interest (ROI)

Compared to edge detection, ROI considers the overall image, as it represents a high level HVS model, however, due to this it can be complex as it requires preparation or analysis to define the ROI. This associated level of complexity is

related to the implementation for ROI within a PVC, as it involves tracking. One application of ROI is video calling, where an oval shape detector is used to identify a person's head and assess the skin colour (Jin and J. Chen, 2009). Then under rate-control, a bit budget is set with a minimum threshold of 60% of bits allocated for the ROI. While this may seem excessive, the application is video calling, where the resolution is low and the relative proportion of the user's head to the screen is large. The results for this PVC solution reflect video calling application as the two test sequences used, Carphone and Foreman are of CIF resolution, both meeting the intended application of video conferencing. The connected nature of ROI is appealing for PVC, however, it requires a complex means to identify and track at the frame level which is unsuitable for the native sub-block level.

2.15 Towards general application use with multiple HVS models

While the above examples discussed specific perceptual techniques of edge detection and ROI to particular applications, they have limited robustness and appeal. The next evolution of PVC solutions is to combine various HVS models together to produce a general multi-HVS model based solution, at the expense of increased complexity. Subsequently, the HVS models chosen may be curtailed or approximated, to balance the increased complexity against the perceptual robustness. This includes using a variety of techniques together, from statistical based HVS models to JND and SSIM algorithms. However, this approach treats the PVC solution as an application using different set of HVS models, than one which has been integrated as one solution. While this is not ideal, it does highlight the technical problems of integrating individual solutions aimed at different levels. Overall, this shows the maturity and ambition of providing an alternative to existing distortion metrics.

2.15.1 Using statistics during video encoding

Video encoding is a very exhaustive process, where candidates are evaluated at different stages. By monitoring these activities, the use of statistics have led to a general PVC solution (Ma et al., 2011). This is based upon monitoring statistical properties for the original pixel array, motion vectors and frame differences. Then

by combining several statistical factors, a rate-quantisation based PVC solution is presented. However, as the video sequences used for training and testing were largely the same, there is doubt that this solution is successful as claimed.

2.15.2 Conditional PVC workflow

In another approach, statistical information along with JND is used to present a perceptual rate-distortion solution (H. Wang, Qian and Liu, 2010). As JND can provide sensitivity of the source and statistical information, such as the rate of change, this can be used to track the perceptual significance of the encoding changes. In this case, mean (μ) and standard deviation (σ) of the JND based difference block was used in selecting block partitioning. The results were limited, yet they showed significant time savings (minimum of 70% encoding time), with little variation in PSNR. However, the results showed fixed QP settings only under a single profile of low delay P. Since JND is a perceptual model and not a measure, the use of tracking changes via statistics allows dynamic thresholding to occur. This means JND is the rate of change relative to the surroundings and not the individual pixel content Using threshold based JND may restrict when to apply PVC, however, the workflow design requires both JND and statistical complexity to be calculated initially. This means that the potential complexity saving benefits of conditional PVC are not fully realised by design, instead by implementation to achieve the encoding time savings.

2.15.3 Combining several HVS models to operate at different levels

The use of multiple perceptual techniques can increase robustness and complexity, however, these can also occur at different stages within the encoder. As HVS models are designed to operate at different stages or levels, combining these would seem ideal, but at the cost of compatibility and complexity. This is demonstrated with a combined HVS model around the JND profile and SSIM (T.-H. Wu, G.-L. Wu and Chien, 2009). For intra frames both luma masking and texture masking HVS models are applied, while in inter frames JND Δluma profile is factored alongside inter masking consideration of motion vector analysis. Other contributing factors for inter QP is the use of skin detection and an averaged pixel

calculation of SSIM across one block. The solution presented can be described to operate at the mode decision level with additional QP adjustment during encoding. While the tests are set up with an initial QP, there is no mention about how one would operate under rate-control where bandwidth restriction would be applied. The results are significant, between 5% and 30% bit-rate reduction for similar perceptual quality, however, this is for CIF resolution video only and timing is not discussed. Overall, using JND with SSIM highlights how different perceptual techniques can complement each other, if they can be integrated successfully.

2.16 Summary of chapter

The hybrid block-based video encoder has been proven to be an efficient, as shown by the adoption by the video coding standards of MPEG 2, H.264/AVC and HEVC. Each generation has increased the choice of candidates per block to enable higher success of block matching and minimise the pixel differences required to be encoded. As the number of candidate choices have increased, the distortion assessment has greater influence on what information is preserved. Being able to assess the pixel differences by their perceptual significance can ensure encoded bitstreams are HVS friendly. However, the video coding standards employ STDMs, which apply uniform cost and do not reflect the non-linear response of the HVS. Applying HVS models in video coding has a limited effect, as they only focus on the original video content. This has resulted in PVC solutions which employ λ_p, to steer quantisation perceptually, however, these do not assess candidates individually. Understanding why and how these PVC solution are applied will be discussed in the critique in Chapter 3 Critique of relevant literature. Finally, objective measure of PSNR or SSIM can provide a guide of image or video quality, however, for a PVC solution these are no substitute for subjective testing, where people evaluate the encoded video performance.

Chapter 3

Critique of relevant literature

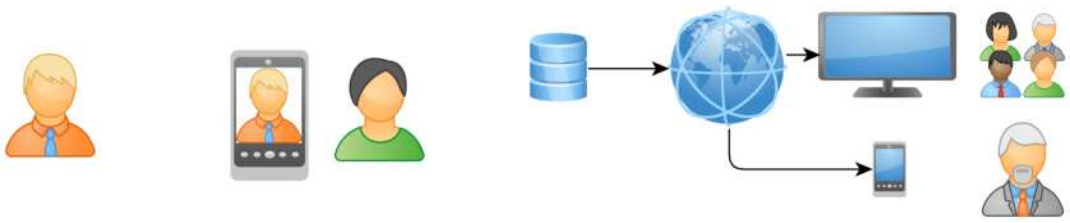
From the previous chapter, the background of video coding standards and developments in PVC were discussed including HEVC and SSIM respectively. This highlighted the motivation for a PVC solution is to make encoded bitstreams HVS friendly and offer a perceptual gain over those based on STDMs. Previous PVC solutions incorporated HVS models, aimed specifically for certain applications, while existing PVC solutions employ λ_p for general applications, however, both suffer issues of high complexity. This chapter will highlight how new applications for video coding are aimed at low powered and/or portable devices, which do not agree with existing PVC solutions that are computationally expensive. Therefore, this critique illustrates why these existing PVC solutions based upon λ_p struggle, explaining the issues of score incompatibility and high complexity associated with the SSIM IQA algrothim. These issues prevent PVC being applied to low powered and/or portable devices where demand is expected to grow. This critique presents a concept of the ideal approach, where a new series of pixel level IQAs are applied across the front-end encoder and the VCL is evaluated with a new tool. This approach is broken down into stages, listed as challenges and presented

as an overview of experiments from which the proposed low complexity in-loop PVC solution can be realised.

3.1 Applications for low complexity PVC

From 'Chapter 2 Background of video coding, PVC and SSIM', HEVC was targeted towards low powered and/or portable devices, where bandwidth and processing constraints can affect user experience. As video coding support is moving towards low complexity devices, the series of applications for which video encoding are required for a human audience can be categorised as either storage or response critical. Storage critical supports greater compression efficiency which increased encoding times. While response critical is where compression efficiency is sacrificed for encoding times. These two categories are shown in Figure 3.1 and Figure 3.2.

These different applications can be explored by the level of support for certain features. Table 3.1 indicates how the different bit-rates associated for each of



(a) Local storage video acquisition and playback - high to medium bit-rates

(b) Television broadcasts and video on demand (VoD) - medium to low bit-rates

Figure 3.1 Storage critical - video coding applications for low powered devices and bit-rates

(a) Video communications - low to very low bit rates

(b) Multimedia wireless sensor networks (or Internet of Things (IoT)) - very low bit-rates

Figure 3.2 Response critical - video coding applications for low powered devices and bit-rates

	Storage critical		**Response critical**	
Parameter	Local storage	Broadcast/VoD	Video comm.	WSN/IoT
Bit-rate	High to Medium	Medium to Low	Low to Very Low	Very Low
Resolution	High	High	Medium	Low
Frame rate	High	Medium	Low	Very Low
Quantisation	Low	Medium	High	Very High

Table 3.1 Different video coding applications for low powered devices and their relative value per parameter

the four major applications led to compromises for resolution, frame-rate and quantisation. For the first two applications of local storage and broadcast or VoD, high/medium bit rate, resolution and frame rate places a demand on resources and complexity of encoding. This encourages the use of non-reference frames which are highly compressed allowing for high levels of efficiencies in video coding at the expense of high complexity. In comparison, the use of resources and complexity are low for the last two applications of video communications and when under wireless sensor networks (WSN) or Internet of Things (IoT). Under these circumstances, a demand for low/very low bit rate can mean high levels of quantisation with reduced resolution and frame rate. This means that for low complexity devices to offer these applications, they must support these range of values per parameters. This is important to understand as any PVC solution must minimise its complexity footprint as not to restrict where it can be applied. In Section 2.13 Integrating HVS model(s) within a PVC solution, existing PVC solutions that employ HVS models increase complexity substantially. Therefore, PVC solutions should be investigated whether they are suitable for low complexity environments.

3.2 Video coding layer (VCL) and need for pixel based PVC

The video coding layer (VCL) contains the final encoding decisions within the bitstream, include prediction candidate choices used during mode decision, motion vectors, along with associated motion compensation and the quantised tranform coefficients (Sullivan et al., 2012). This means that the coding structure containing the signalling and picture information is stored in the VCL. These coding choices

occur during the front-end stages of prediction, mode decision and rate-control, where distortion and activity assessment are applied at the pixel domain. While the later stages of transform, quantisation and entropy encoding represent the back-end stages, operate in the frequency domain, which suppress coefficients of residue and not influence the signalling choices made by assessment (Y. Kim et al., 2012). Ideally, PVC should occur at the earliest possible point, which is at the front-end stages, in the pixel domain, so each candidates can be evaluated individually. Also, for frequency based PVC encoders, frequency perceptual assessment will be limited to differences unless the original and/or reconstructed are also transformed, which would mean additional processing. In all, pixel based PVC can allow perceptual assessment using original pixels or calculate reconstructed pixel values with just additions or subtractions. Prior to IQAs, perceptual models operating on the frame level, such as JND, would be used to steer the back-end encoding stages (Yang et al., 2005). This approach does highlight the need that pixel based PVC is beneficial for applying edge detection, as local edge detection is fundermental aspect of the HVS (Elder and Zucker, 1998). Overall, such approaches combines both domains, however, it leaves open the ability of candidate selection by perceptual assessment. Therefore, pixel based PVC will affect the candidate selection for the VCL, allowing great diversity in bit distribution while maintaining perceptual integrity.

3.3 Lack of a low complexity PVC solutions

While PVC solutions are shown to be effective at identifying perceptual redundancy, however, the management of complexity can be gauged by how frequently and the type of operation used (Y.-H. Huang, Ou, Su et al., 2010; J. Kim, Bae and M. Kim, 2015; Yeo, H. L. Tan and Y. H. Tan, 2013). This means that rather than complexity, overall encoding timing is shown, which allows masking the high complexity design of these existing PVC solutions. When timings are considered for HD video sequences at 1080p resolution, the results show increases of +10% and +25% for low delay P and random access respectively (J. Kim, Bae and M. Kim, 2015). Their findings are competitive compared to other block based PVC solutions, however, these results are high compared to the reference encoder, suggesting a sub-block PVC encoder will need to be complexity aware. This need for a low complexity PVC solution is recognised by broadcast industry and academia through

joint collaborations such as PROVISION and THIRA (Provision-itn.eu, 2015; Thira, 2015). Equally, the projects like COGNITUS acknowledge that the Internet has allowed users generated content and broadcasting to converge, and thus are working on producing low complexity HEVC encoders (Codec, 2016; Cognitus, 2016). Overall, existing PVC solutions are focused on perceptual coding gain to improve bandwidth efficiency, when the trend is for low powered devices which require low complexity solutions.

3.4 Existing PVC solutions

The early generation of PVC solutions were limited to application specific as discussed in Section 2.14 Application specific HVS models, which led to the current form of λ_p based PVC solution as indicated in Section 2.12.3 Suitability of SSIM in video coding. This section will explain how λ_p based PVC solutions operate, why they are approximate models and how come they are still computationally expensive. These existing λ_p based PVC solutions have incorporated perceptual coding techniques outside the native sub-block level of the encoder because the high complexity associated to scale the incompatible SSIM scores with STDMs. This approach results in a relatively static PVC, as the model updates less frequently than the candidates at sub-block level. Newer forms of λ_p based PVC solutions have led to more frequent PVC model updates, by splitting the IQA equation or by making assumptions, however, they do not provide an in-loop PVC solution.

3.4.1 Existing approach to perceptual lambda

The existing approach to PVC is to applying perceptual distortion assessment outside of the sub-block level workflow, by way of a model to rescale quantisation. This is shown in Figure 3.3, whereby the quantisation value following rate-control is intercepted and replaced with one based upon a perceptual R-D curve. By producing a perceptual R-D curve, this model can be used for various blocks within a frame or for frames within a GOP. This type of solution is common, as the IQA is complex to run and the translation of a perceptual score to a compatible STDM equivalent is also highly complex. The weakness of this solution, is that while it operates outside of the sub-block level, which keeps the overall complexity low, the perceptual calculation is not applied individually for each candidate. This can

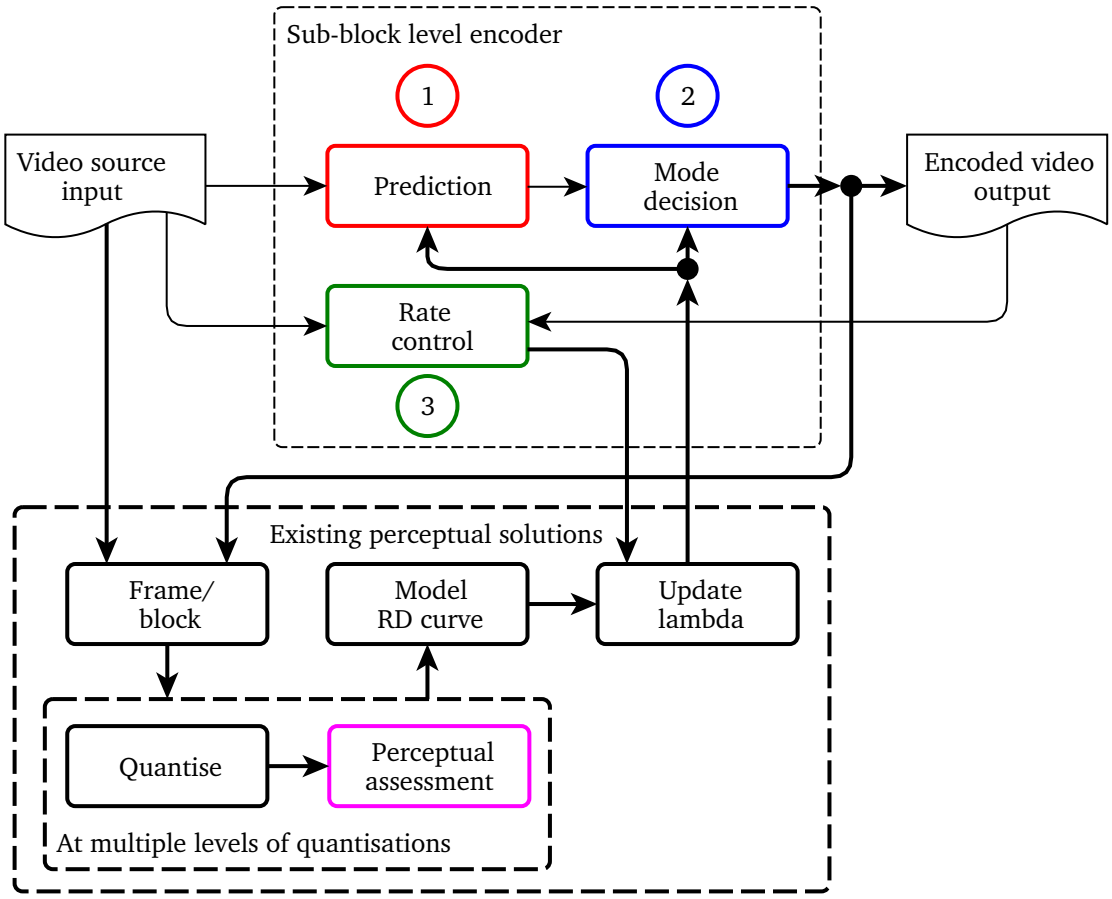


Figure 3.3 Existing approach to perceptual video coding

be illustrated with figure 3.4 where perceptual λ model is based upon a single perceptual R-D curve which may not be suitable for other candidates within the sub-block, block or frame. The ideal perceptual R-D curve varies according to the candidate, while existing PVC solutions apply a perceptual λ based on a fixed model with tangents along the perceptual R-D curve. This approach saves in terms of complexity, however, it is not optimal at ensuring an HVS friendly encoding.

3.4.2 Perceptually rescaled quantisation

The perceptually rescaled quantisation treats the block and its sub-blocks the same. This means the encoder is inflexible to change per candidate, which restricts the perceptual R-D to a single curve. As quantisation is rescaled, the point along the R-D curve which is selected is changed, where quantisation represents the tangent of the R-D curve as shown in figure 3.4. In order to rescale quantisation, existing

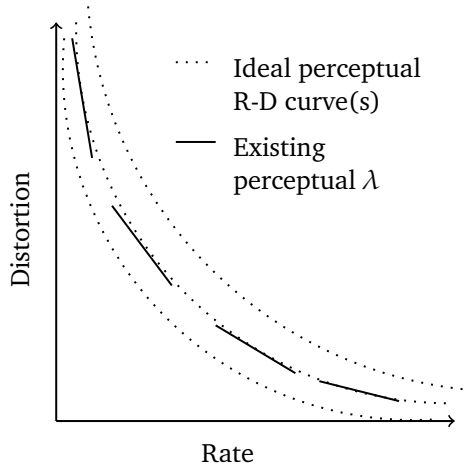


Figure 3.4 Ideal perceptual curve against existing perceptual λ

PVC solutions produce their own form of R-D curve model based upon an IQA for measuring distortion, which involves the frame being encoded under different levels of quantisation (Y.-H. Huang, Ou, Su et al., 2010). This raises the issue of scaling an IQA score to be compatible with existing forms of assessment, which means perceptual rescaling of quantisation. These processes can be illustrated with Figure 3.5, where the traditional R-D curve is mapped against the R-Dp (perceptual) model, to produce the rescaled quantisation value. With this restricted R-Dp curve response, this may translate to the same distortion score. However, this rescaling of quantisation is a computationally costly process, and so R-Dp is predetermined at the GOP to keep complexity low.

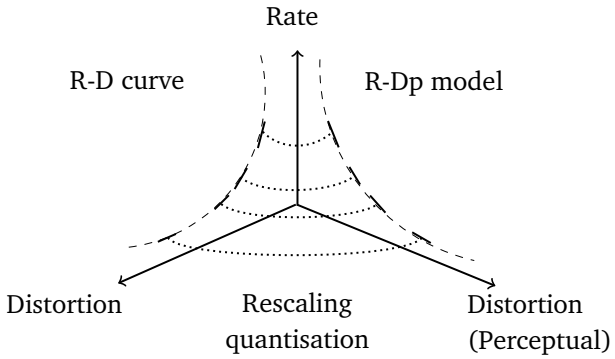


Figure 3.5 Mapping R-D (STDM) curve against a R-Dp (perceptual) model to rescale quantisation

3.4.3 Using quantisation rescaling to lower overall complexity

Prior to using λ_p, attempts at bringing an IQA in video coding would incur significant overhead. Perceptual quantisations tackles this issue by splitting the processing of the perceptual model generation and the quantisation rescaling into two, which lowers the overall complexity. However, as the model updates less frequently, λ_p may not be appropriate. To address this issue, other forms of tracking perceptual changes have been introduced to update the perceptual accuracy. This has meant linking perceptual related components used to calculate perceptual assessment with less complex STDM calculations, to only apply it to certain block sizes, or on a conditional basis (Dai et al., 2014; Qi et al., 2013; Yeo, H. L. Tan and Y. H. Tan, 2013; Yuan et al., 2013). While these allow more frequent model updates down to the block level, they do not update on the basis of sub-block candidates individually. This means they are unable to capitalise upon opportunities to perceptually assess individually, which could further improve perceptual gain or integrity.

3.4.4 Limitations of perceptual quantisation

The common approach of existing PVC solutions places complexity associated with an IQA outside the sub-block level workflow. As a consequence, λ_p keeps the overall complexity low, at the expense of limiting the frequency of calls to an IQA. However, λ_p is applied based upon mode decision, the computation expense is greatest on mapping SSIM to SSE as part of the quantisation rescaling (Y.-H. Huang, Ou, Su et al., 2010). The incompatible scores of SSIM mean complex operations are used as when curve fitting R-Dp to R-D, which dramatically increases complexity. This has led to calls for a compatible IQA, one which can be used in-loop, during the respective sub-block stages for a native PVC solution (Su et al., 2012; F. Zhang and Bull, 2015).

3.4.5 Subjective performance of PVC solutions

Existing PVC solutions are able to provide credibility to their proposed solutions by demonstrating subjective testing results. One method is the double-stimulus continuous quality scale (DSCQS) based upon ITU-R BT.500 (ITU-int, 2012). This methods was used to show video content of different resolution, comparing the

proposed against two alternatives (J. Kim, Bae and M. Kim, 2015). The set-up involved showing nine videos at four different QP settings to each subject, however, the number of participants were not revealed. While this test set-up tried to adhere to ITU-R BT.500, as different video resolutions were used, this meant that viewing distance should be changed, yet this was not mentioned (Demers, 2016). Equally, the use of fixed QPs is not representative of today's applications where bandwidths are likely to govern transmission or recordings.

3.5 SSIM in PVC

The current generation of PVC solutions adopt SSIM as the IQA of choice to produce R-D perceptual (R-Dp) curve than to perform perceptual distortion assessment on candidates. This is because of the limitations of SSIM which restrict its use as a direct replacement for STDMs. These limitations are centred around how SSIM scores are relative than absolute, which makes it difficult to compare assessments between block-sizes. To understand this, consider the RDO process during mode decision where several smaller sub-blocks are compared to a large (sub) block. For STDMs are distortion metrics the scores can be accumulated which allows comparing smaller sub-blocks with larger (sub) block(s). Unfortunately, SSIM is an index (bounded by ± 1) meaning that values can not be accumulated like in STDMs. Instead, λ_p based PVC solutions are applied to an R-Dp curve retrospectively at the block level, which ensures the same spatial size is assessed. This restricts SSIM to the block level, relying on STDMs to individually assess each sub-block candidate. Overcoming this obstacle requires addressing the complexity of scaling a SSIM score to an STDM compatible score.

SSIM is accepted as an alternative to PSNR for assessing image coding and video coding. However, when integrating SSIM to the video encoder workflow this presents challenges in terms of compatibility and complexity. Also, as SSIM is being understood, its perceptual effectiveness is being questioned with alternative perceptual techniques which have increased complexity.

3.5.1 Compatiblity of SSIM score with STDMs

Due to the manner in which SSIM is calculated, it means SSIM can not be used as a direct replacement for STDMs. Having compatibility is important, to ensure the

rest of the encoder can operate with minimal changes. This means that the SSIM score must relate to an existing STDM score. As part of existing PVC λ_p solutions, the SSIM score must be scaled so it may be compatible with existing STDMs. This rescaling of quantisations by PVC solutions requires mapping SSIM to an STDM score by a process of conversion which is highly complex.

3.5.2 SSIM vs STDM scores

The SSIM algorithm presents a score which is of continuous form bounded by ± 1 as it is part of an index, irrespective of window/block size. In comparison, STDM is discrete, where the score is whole integers score which is bounded only by the bit-depth and block size. SSIM is a rating scheme against the reference, yet the differnces against the maximum score correspond to relative than absolute pixel differences. This means that the relative scoring of SSIM can be called 'ordinal numbering system' as the distance between scores is not representative of the differences. While for STDMs, the scores are calculated to the actual differences in pixels, which allows it be classed a 'ratio numbering system', especially as zero means no pixel differences. These different number types mean SSIM is not natively compatible with STDMs. This incompatibility also exists when comparing different sized sub-blocks based upon different SSIM window sizes, especially during RDO.

3.5.3 Triangle equality rule and the geodesic triangle

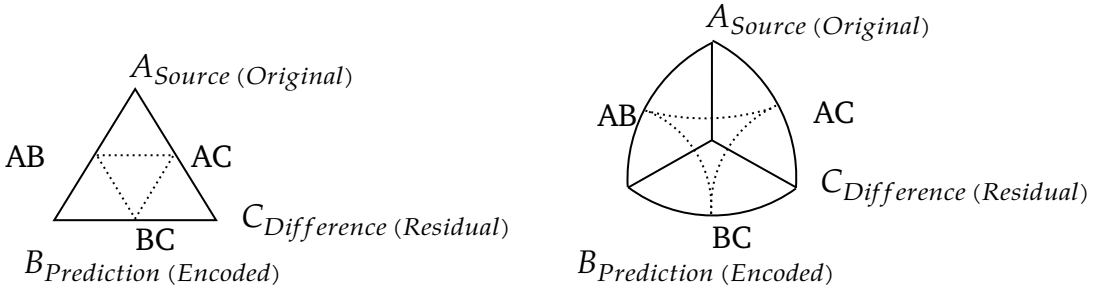
The above highlighted that the issue of incompatibility lies between SSIM and STDM, however, this is because SSIM does not share the same numerical properties as the STDMs. From a mathematical perspective, this relates to the fact that SSIM does not satisfy the triangle equality rule ($\triangleq$). The $\triangleq$ is fundermental to define metric based upon a uniform distance numbering system, which perceptual measures like SSIM do not support. A uniform distance numbering system allows the addition and subtraction of distances to correspond to the sum and remainder as measured. The $\triangleq$ is defined as where the sum of two sides is greater than the third, as shown in Figure 3.6a. Existing STDMs satisfy this condition as the differences that relate directly to the remaining score. While for SSIM, the score is bounded by an upper and lower limit irrespective of block size or bit-depth, which means $\triangleq$ is not satisfied. However, there is a concept that SSIM should be described as

a geodesic triangle equality, see Figure 3.6b. This involves scaling it to be $\triangleq$ with a single function which is extremely processor intensive and difficult to resolve (Richter, 2011).

Modelling SSIM as a geodesic triangle was shown to be resolved when the constants C_1 and C_2 was set to zero (Brunet et al., 2012). While successful, its applications are limited until a general solution exists, where the constants are non-zero to provide a stable function. As part of producing a mathematical understanding, a proposed language was defined to represent the 'visual distance' in SSIM's geodesic (curved) space (Richter, 2013). The work was modelled on Taylor series which helps eliminates the first term and constant so the model can be generalised. This means that for SSIM to be a distortion metric remains unresolved and is not worth pursuing due to increases in computation complexity. Therefore, the existing approach of using SSIM to model R-Dp curves outside the native encoder workflow and STDMs within the native sub-block level remains popular. This existing approach makes a SSIM a pseudo-metric score that is compatible with STDMs.

3.5.4 Addressing SSIM complexity

As SSIM involves statistical moments, it uses a sliding window as part of its calculation, which is far more intensive than single pixel calculation for an STDM. Also, making SSIM into a pseudo-metric compatible score for the purposes of video coding will increase complexity, as another layer must be introduced. This means



(a) Triangle Equality (b) Geodesic Triangle Equality

Figure 3.6 Richter's geodesic triangle equality for SSIM vs the triangle equality rule for distortion metrics

adapting SSIM for video coding brings double layered complexity barrier which far exceeds the STDM complexity envelope.

3.5.5 Integrating SSIM into video coding

In terms of integrating SSIM into video coding, the assessment of motion compensation is a complexity sensitive process. A variant of SSIM was introduced, where alongside spatial consideration, temporal coordinate was factored when calculated (Moorthy and A. Bovik, 2009). The complexity was discussed to be dramatically reduced, however, performance figures were sparse, and the proposed solution operated at the frame level using a fixed 8x8 block. Another approach where a technique from the field of computer vision, using an integral of a processed image to speed up the calculation of moments within SSIM (M.-J. Chen and A. Bovik, 2010). Unfortunately there is a reliance of tracking neighbouring blocks which in hybrid-block video coding like HEVC may be difficult to manage when tiles are implemented (Misra et al., 2013). Upon inspection, both approaches are based upon the same simplified version of SSIM, where it is assumed that $\bar{x} = \bar{y} = 128$ (Rouse and Hemami, 2008). This is not a fair assumption because at the prediction stage the candidates can vary a lot especially because of the content which immediately rules out these fast SSIM solutions. Overall, this means that issue of computational complexity for SSIM remain and this must be addressed for a PVC solution to be adopted by others (Chandler, 2013).

3.5.6 Applying a rolling SSIM calculation

Another means to address the complexity for SSIM calculation is to operate on the basis of monitoring the changes and updating the SSIM score accordingly. This is applied as part of a PVC solution, where during RDO in the mode decision stage, blocks changes are tracked allowing SSIM score to be updated from an STDM (Dai et al., 2014; Yeo, H. L. Tan and Y. H. Tan, 2013). This does allow complexity of SSIM to be aligned with STDMs which has a lower complexity envelope. However, these PVC solutions do not have access to the reconstructed block during RDO. This approach risks relying upon existing design limitations of the encoder, as typically the reconstructed image is not available during encoding. Despite, this it is possible to calculate the reconstructed image on the $\triangleq$ principle. Another issue

is that covariance must be estimated, and it is assumed that the means for both the original and reconstructed are identical. While this lowers the complexity for calculating SSIM, the level of assumptions applied mean it risks not being fully adaptive to the content.

3.5.7 Complexity for scaling SSIM scores to STDM compatible

While complexity for SSIM remains, another layer of complexity is required to scale SSIM scores to be STDM compatible. This additional layer is usually applied slightly differently depending upon SSIM implementation, however, typically it involves complex transform. The transform applied is non-linear, involving logarithmic and exponential operations, which is more complex than the SSIM algorithm itself (Y.-H. Huang, Ou, Su et al., 2010; Yeo, H. L. Tan and Y. H. Tan, 2013). In turn, this allows a scaling factor to be applied as part of a perceptual based calculation for a Lagrange multiplier, λ. Consequently, to minimise the overall encoder complexity, the frequency for re-scaling λ may be conducted via a model which limits the operations to once per GOP or frame (Y.-H. Huang, Ou, Su et al., 2010; Yeo, H. L. Tan and Y. H. Tan, 2013). This is not ideal as while the overall complexity is managed for the PVC solution, its implementation limits how dynamic it adapts to candidates and content.

3.5.8 Effectiveness of SSIM

It has been shown that when SSIM is evaluated against PSNR using MOS, SSIM is favoured over PSNR (Z. Wang et al., 2004). When examining SSIM under specific types of noises, it is possible for SSIM to award a higher score to a heavily burred image compared to a noisy image (Brooks, X. Zhao and T. Pappas, 2008). The poor performance of SSIM in these conditions is due to the distorted image having a similar mean and variance despite inherent noise being present (Zujovic, T. Pappas and Neuhoff, 2013). The problem lies in the development of an IQA, where availability of grounded truth images are limited due to the amount of time involved to produce them (Chandler, 2013). Images and video databases do exist to evaluate IQAs, however, the grounded truth process has not been undertaken. This suggests that IQAs like SSIM require additional measures to mitigate against false positives when scoring distortion.

3.6 Need for low complexity PVC and visualising the VCL

Video encoders like HEVC are being designed around portable and low powered devices, as presented in Section 3.1 Applications for low complexity PVC. However, existing PVC solutions are based upon λ_p, which must produce a model and then rescale quantisation, operate outside the native sub-block level, as shown in Section 3.4 Existing PVC solutions. The reason for this re-scaling is explained as SSIM is incompatible with STDMs, as SSIM does not fulfil the $\triangleq$. Consequently, this means that existing λ_p based PVC solutions have high levels of complexity by design, making them unsuitable for low powered devices and/or portable devices. This issue has been raised by others, stipulating that PVC should occur natively within the sub-block level.

Another issue that is critical in the development is to verify that the IQA is applied in the correct regions and to validate changes in the encoded bitstream at the correct regions. For video coding, the IQA development is complexity critical, therefore, using the encoding environment to development complexity aware IQA requires a visual tool verify and validate implementation.

3.6.1 Calls for in-loop IQA at the sub-block level

While existing PVC solutions have shown that IQA is possible, the challenges faced is to lower the IQA complexity to make them competitive to STDMs (Chandler, 2013). This means producing a low complexity PVC solution, which can extend benefits to applications in medical imaging, restoration/de-noising and virtual reality (VR) (Reinhard et al., 2013). It is important to understand that encoding is crucial for modelling and rendering of computer graphics images (CGI). In these environments, IQA presents a means to reduce the intensive rendering times and morphing operations which can lead to energy savings as the perceptually optimal point may state that the distortion is perceptually insignificant (Greenberg et al., 1997; Schödl et al., 2000). Yet, while IQAs are designed for image coding environments than for operating within in-loop video coding environments, the issues of complexity and compatibility remain (Su et al., 2012; F. Zhang and Bull, 2015). This issue is compounded by the increasing amount of video data consumed

and produced using low powered devices. Operating within this low complexity envelope, video coding requires 'codec friendly' design techniques (Su et al., 2012).

3.6.2 Need to visually evaluate performance using the VCL

Any proposed PVC or the development of an IQA must be evaluated to measure it performance. The typical the success of PVC solution is by incorporating subjecting testing, this gives credibility to a PVC solution aimed at the HVS. There are standards developed which can guide the set-up of a subjective test, are described in Section 2.6 Subjective testing. However, the development of an IQA can not undergo continual subjective evaluation due to availability of participants and resources. Neither, can the sole evaluation be done by objective testing of PSNR and SSIM where numerical results are presented. This means that a visual representation of existing STDMs, objective measures and proposed IQA(s) is required to illustrate how different regions are affected. Typically, development of IQA for video coding has largely focused upon adapting perceptual image assessments designed under mathematical software. Calls have been made that IQA for video coding should have an assessment suited for this environment and the use of mathematical tools may not be best for its evaluation (Chandler, 2013). Adapting an algorithm for video coding can present its own set of implementation related challenges. Being able to operating in a similar environment video coding during development can present a closer integration of the final solution.

3.7 Ideal approach to low complexity PVC and to visualise the VCL

In this section a low complexity in-loop PVC concept will be presented, however, for this to be realised a new generation of IQAs will be required, called pixel based IQA. This new pixel based IQA will need to work in conjunction with STDMs to provide a low complexity in-loop PVC solution. Due to this integrated approach traditional means of IQA development are not suitable, instead a visual tool is required to illustrate the VCL and simulate the proposed IQA.

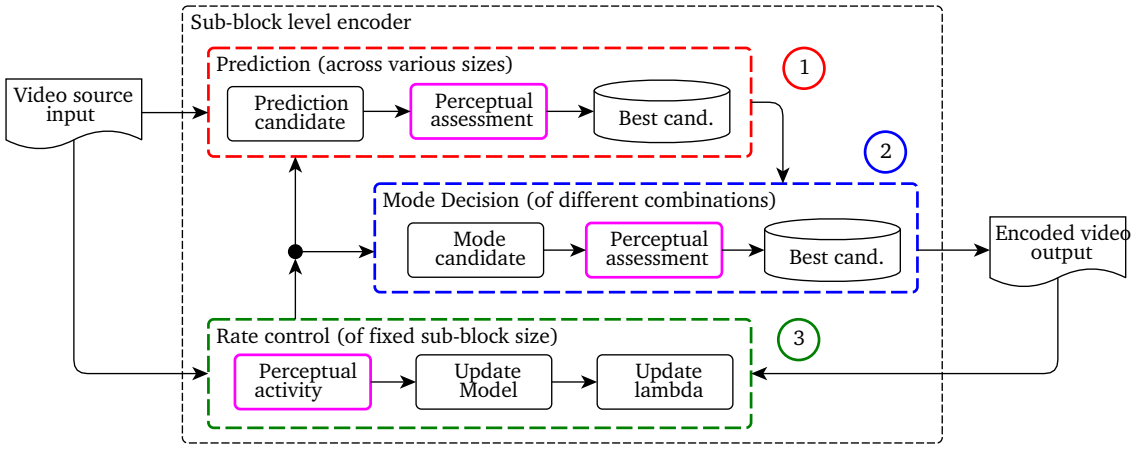


Figure 3.7 Ideal approach: sub-block level PVC

3.7.1 Ideal approach: sub-block level PVC

The existing approach to PVC with λ_p is unable to operate at the native sub-block level. As presented in Figure 3.7, the ideal solution is where each of the three stages, prediction, mode decision and rate-control of the sub-block level encoder workflow has a native IQA. This means that the for prediction and mode decision existing STDM should be replaced with a perceptual equivalent. For rate-control the ideal should include perceptual activity assessment as it will allocate bits according to its perceptually significance. In an ideal PVC solution, this will result in perceptually redistributing bits to retain perceptual integrity of the video sequence.

3.7.2 Ideal IQA: Single pixel-based IQA

A common theme that was shown in Section 2.8 Existing perceptual techniques was that HVS models or perceptual techniques were shown to be highly complex. This is because HVS models or perceptual techniques require multiple pixels locations from which to calculate a score. SSIM is no exception as it uses a sliding window to calculate the relative differences to its surrounding pixels, which compared to STDMs that calculate pixel differences pixel only. This design approach of pixel based IQA will dramatically reduce the complexity burden and allow sub-block level assessment. However, this different approach needs a tool to verify its perceptual performance suitable for a low complexity sensitive environment.

3.7.3 New tool to visualise the VCL and simulate IQAs

Perceptual assessment within PVCs play a significant role in being able to influence the video encoder. This can result in the redistribution of block sizes, choice of candidates and bit allocation, which can change to the VCL of the video sequence frame. The evaluation of these PVC solutions have traditionally being indirect, either via objective measures or subjective testing. Judging this directly means accessing the bitstream file and producing heat maps to illustrate the signalling values. Equally, if the original source video sequence is read then IQAs or STDMs can be simulated, illustrating how different regions retain information across the video frame. Existing tools which analyse the VCL are usually commercial and restrict adaptation of the code base. While image quality tools use decoded video where meta information such as the signalling and partitioning are not available. Finally, a visual tool operating with an encoded bitstream can be applicable to any compatible bitsteam, which enables comparisons to be take place without the need to store large decoded video sequences.

3.8 Challenges in existing research

In the previous sections, a critique was presented for a low complexity in-loop PVC solution and a tool by which to visually evaluate performance. This conclusion was supported by others, with a call for an IQA that is compatible with existing distortion metrics and offers low complexity methods to enable in-loop calculations. Currently, PVC at the sub-block level is not understood, meaning that this research provides an opportunity to explore this area before approaching low complexity IQA.

A single pixel based IQA could be more compatible with STDMs and enable sub-block level PVC, however, existing IQAs at the sub-block level is not understood. Consequently, the relationship of STDMs and SSIM at the sub-block level must be first explored and understood, before any attempt at single pixel based IQA(s). This first stage is important to identify features that have not been identified and can be designed into a single pixel based IQA(s). Also, since STDMs operate at different norm spaces, there is a likelihood that multiple pixel based IQAs will need to be developed.

This section aims to address these issues and present them as a series of research challenges from which these will be addressed in the form of experiments. As research into the sub-block level is uncharted, an initial experiment will need to consider the behaviour of SSIM at the sub-block level. The subsequent experiment will address the complexity of mapping perceptual to STDM. Finally, this allows the development of low complexity in-loop IQAs specific to each front-end encoder stages; prediction, mode-decision and rate-control. However, this development requires a new visual tool, demonstrating the effects on the partition structure and meta information within the VCL. In all, these challenges are around reducing the complexity footprint of a PVC solution, and thus making it viable for low powered devices in the same way as the existing encoder.

3.8.1 Exploring the sub-block level with existing SSIM IQA

Studies comparing perceptual distortion assessment and STDM have shown a non-linear relationship (Horé and Ziou, 2013). These have reinforced the idea that existing STDM do not reflect the HVS, propelling the development of PVC solutions. Existing studies into perceptual assessment techniques with STDM have focused on distortion scores taken from rendered frames or blocks. These existing methods are outside the native sub-block candidate selection process of the encoder workflow, which apply the same R-Dp curve. An IQA can have a different response curve based upon the content of original and reconstructed sub-blocks. Also, the curve can be different due the level of quantisation. Together, this means that perceptual and STDM cover a region of space based, with a range of values. By conducting an observational study, within the sub-block encoder workflow, the potential volume

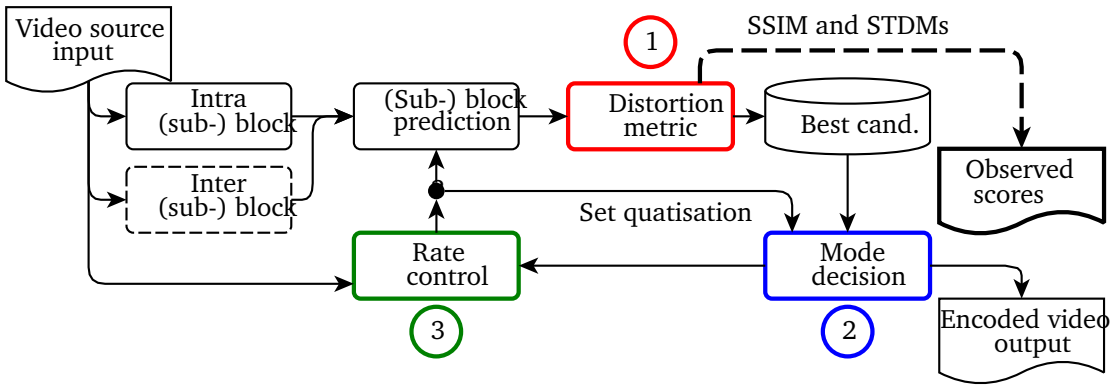


Figure 3.8 Observational study of SSIM and STDM scores at sub-block level

of observations can illustrate this region of space. Furthermore, the sub-block level has not been examined and this will provide evidence in terms of SSIM and STDMs. Overall, this first research challenge is designed to justify exploring the sub-block level and is illustrated in Figure 3.8, where observation samples are gathered from the prediction stage. As this occurs at the prediction level, both intra and inter sub-blocks will be assessed, allowing for a range of values to be recoded for SSIM and STDMs. In addition, as these will be paired observations, this will allow analysis examining for any pattern between SSIM and STDMs scores.

3.8.2 Low complexity scaling of SSIM to STDM score

Scaling SSIM to a STDM scale ensures compatibility with the assessment in mode decision where RDO takes place. As the scaling of SSIM represents a significant proportion of complexity, a more intimate relationship between SSIM and scaling must be explored. This means discovering new relationships between components of SSIM with that of STDMs, in order to lower the complexity for scaling SSIM. This has two benefits, firstly, each sub-block candidate can have a perceptual score adapted for different Luma values. Secondly, finding an association between SSIM and STDMs can lower the scaling complexity, reducing one of the initial major challenges. Effectively, this challenge aims to implement SSIM within the front-end stage of the encoder as part of a PVC solution. Unlike existing PVC solution which present them as λ_p, this challenge is about introducing SSIM assessment at the sub-block level. This is shown in Figure 3.9, where under 'stage 1', the distortion metric has been replaced with SSIM assessment. This means that a prototype, proof of concept will be shown, where the best candidates during 'stage 2', in mode decision

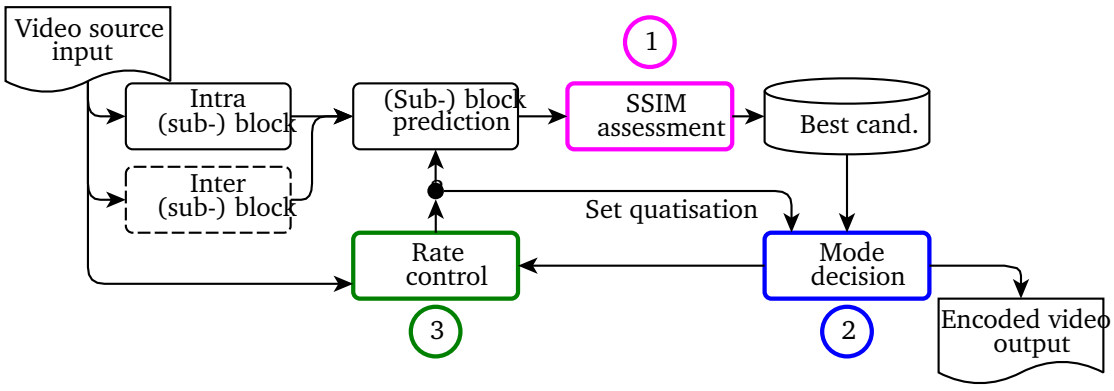


Figure 3.9 Encoder with SSIM distortion assessment at the sub-block level

are encoded. Overall, being able to apply SSIM at the native sub-block level, with a low complexity scaling will present a novel approach for making SSIM compatible with STDMs.

3.8.3 Low complexity in-loop PVC

These initial experiments are largely dependent on finding a means to fit SSIM into the encoder's sub-block workflow. While this was limited to SSIM at the prediction stage, ideally, perceptual assessment should occur at every front-end stage. Considering that SSIM is complex compared to STDMs this is not attractive for low complexity PVC. Instead, new IQAs are required which are low complexity by design and suitable for the respective front-end encoder stages. This is shown in Figure 3.10, where each of the three stages have sub-block level IQA or activity. As this development aims to promote pixel-based IQAs specific to each front-end stage, it will mean developing several IQAs and producing perceptual integrity tests designed to evaluate when best to apply them. This means that these new IQAs would work conditionally alongside the STDMs as a means to steer the encoder to favouring one candidate over another.

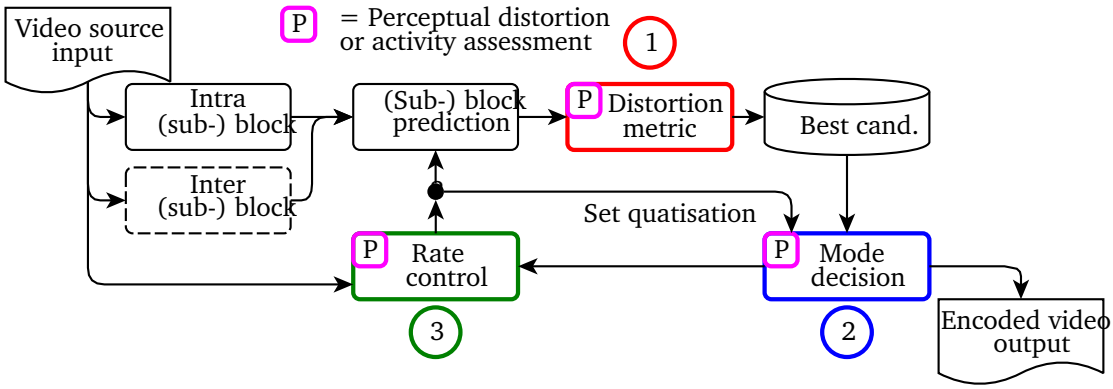


Figure 3.10 Low complexity in-loop PVC

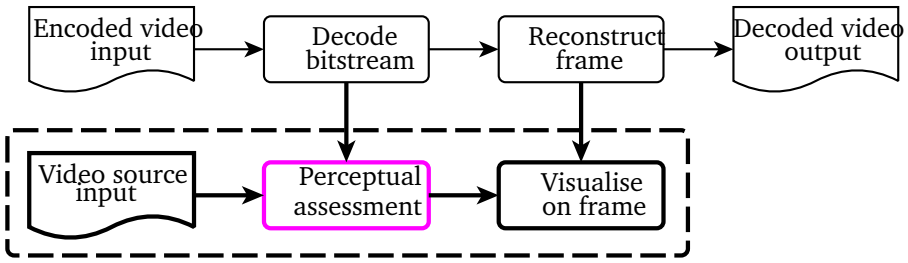


Figure 3.11 Modified decoder with IQA visualised on frame

3.8.4 Visualising VCL changes on bitstream

The need to design low complexity IQAs means a new tool is required to see the effects of implementing new PVC solutions on video coding. The literature review indicated that an IQA should be called selectively, to ensure that overall complexity remains low. This calls for a combination of perceptual related techniques than a single IQA, as effective and efficient use of resources. This means that the current manner of a single algorithm solution devised and modelled on mathematical software is not suitable. Also, traditional methods to evaluate IQAs are seen as cumbersome as they are time and resource consuming (Su et al., 2012). Consequently, there is a need to produce a new visual tool to assist the development of these new IQAs. The decoder can allow the extraction of signalling information and simulate the assessment without affecting the encoding process. Figure 3.11 illustrates the modification to the decoder such that perceptual and signalling can be visualised on the video frame. This tool will be significant in that it can be applied to encoded bitstreams including those of the reference encoder, thus allowing visual comparisons to be made. In all, a visual tool can provide an improved understanding of perceptual and existing forms of assessment and its effects.

3.8.5 Subjective testing

An aspect of testing which affects the validation of a PVC solution is the use of subjective testing. Integrating perceptual assessment inside a PVC solution should be validated to show perceptual integrity is maintained during the video encoding. This is crucial as a PVC solution is aimed at the HVS and subjective assessment is the best possible means by which to evaluate performance. While existing standards do provide sufficient guidelines to set-up testing, they can be liable to biases in the order of original and proposed encoded video. Therefore, it is proposed that while the order of video sequences remains unchanged, the respective order of original and proposed encoded video sequence is randomised. This minimises the risk of bias for first preference and provides greater credibility in the findings for the subjective testing.

3.9 Summary of chapter

This chapter has described that existing λ_p based PVC solutions are not suitable for the new intended application of HEVC. Underlying this is the incompatibility of SSIM with STDMs scores because SSIM does not support the $\triangleq$, meaning that λ_p based PVC solutions operate outside the native sub-block level, which makes them are complex by design. Consequently, calls have been made for a low complexity in-loop PVC solution, allowing PVC to occur on low powered and/or portable devices. This chapter described the call for a low complexity in-loop PVC, as an ideal solution, yet stipulates that these new form of IQAs be pixel based IQAs which integrate with STDMs. As such, to achieve this ideal PVC solution an exploration of SSIM and STDMs relationship at the sub-block level is required before harnessing this understanding to developing the proposed pixel based IQAs. However, the development of these proposed pixel based IQAs involve integration with STDMs they will need for a new visual tool which can simulate when these IQAs and monitor their effect on the VCL. Such a tool would aid in evaluating the performance of the proposed pixel based IQAs.

Chapter 4

SSIM-STDMs relationship at the sub-block level

From the literature review SSIM based PVC solutions are striving towards the sub-block level by attempting to reduce the associated complexity in scaling SSIM scores. The sub-block level represents the native point for the encoder, where prediction candidates are assessed individually and that the mode decision draws upon to represent a block. Existing SSIM based PVC do not cover the sub-block level, because this requires scaling SSIM scores to be compatible with STDMs scores used during the predication stage. Reconciling SSIM with STDMs is vital in order to provide SSIM based PVC at the sub-block level, which has the greatest potential to produce a HVS friendly encoding. However, existing understanding of the SSIM and STDMs relationship has resulted in different interpretations, either as a non-linear response or as something more complex which is difficult to model (Brunet et al., 2012; Horé and Ziou, 2013).

It is important to understand the SSIM-STDMs relationship accurately, to construct a representative SSIM based PVC model. An accurate sub-block level model of the SSIM and STDM relationship can result in prediction candidates

being assessed by SSIM with STDMs compatible scores, meaning that each sub-block candidate is assessed individually. However, at the prediction stage it is important to also have a model which is respective of any additional complexity that is introduced. Existing SSIM based PVC solutions employ high levels of complexity to scale SSIM scores to be STDM compatible. This makes SSIM based PVC solutions unattractive for low powered and/or portable applications, meaning that it must be done with the intent to produce a low complexity model to scale SSIM scores to be STDM compatible. Therefore, this chapter is where the SSIM and STDMs relationship at the sub-block level is observed, examined and then modelled.

4.1 Related findings

The role of STDMs is to establish the candidate with the minimum distortion by assessing differences uniformly. While as explained in 'Section 2.12 Perceptual assessment using SSIM', IQA's like SSIM, calculate non-uniformly as they consider relative changes based upon statistics. These two different approaches can result in dissimilar types of scores for the same differences. This issue can be better understood by evaluating a set of sub-blocks triplets (original, prediction and differences) by the $\triangleq$, which will illustrate the challenge of using SSIM as an assessment. It is possible to illustrate this with a few examples, however, being able to apply SSIM at the sub-block level has its own technical challenges, which relate to having access to the original pixel values during the calculation of SSIM. Overall, each of these points will be discussed in this section, illustrating the situation in bringing SSIM to the sub-block level.

4.1.1 Two different approaches to assessment

In STDMs, distortion costs are accumulated uniformly per pixel, this differs from Weber's law, of tolerating a level of relative change that is perceptually undetectable to the HVS (Weber, 1864; H. Wu and Rao, 2005). STDMs seek pixel exactness against the original source rather than the relative lighting changes based upon the perceptual sensitivity. Consequently, there is a potential for bit savings or bit-redistribution by seeking perceptual redundancy compared to using STDMs. For both these approaches of distortion assessment, STDM and IQA are illustrated in Figure 4.1 and Figure 4.2 respectively. In Figure 4.1, differences are aggregated

independent of the original luma score. This means that during higher levels of quantisation or where the source material is of lower image quality, perceptual clues to distinguish between noise and information are at risk. While when an IQA is used to measure distortion perceptually, the score encompasses the neighbouring pixels to consider to produce a weighted result. as shown in Figure 4.2. This approach does mean that it is less likely to be affected by small changes, as they are averaged out across the sub-block array. However, this does means that when all candidate choices have some degree of errors, it is susceptible to promote those that share the same statistical properties, when in fact the differences are substantial (Fei et al., 2012).

4.1.2 Different types of assessment dissimilar scores

The need to resolve perceptual scores to be compatible with non-perceptual is to minimise the need to adjust other components along the encoder workflow. Typically, this has meant conducting observational or mathematical studies to resolve perceptual scores in STDM space (Bhat, Richardson and Kannangara, 2010; Brunet et al., 2012; Y.-H. Huang, Ou and H. Chen, 2010). This has been limited to frame and block level paired observations, yet investigations into the sub-block level have been overlooked. One immediate benefit of operating at the sub-block level is that variety of observations can be gathered, which allows a comprehensive coverage of the shared distortion metric space. This is crucial as it is known that an IQA is designed to consider original pixel values when evaluating the differences which can influence an IQA score. This can be demonstrated in Figure 4.3, where the range of differences can affect the SSE scores, in comparison, to calculate the SSIM scores, the luma sensitivity for the original is also used. Looking at Figure 4.3a

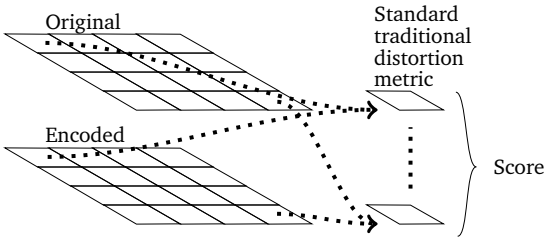


Figure 4.1 Standard traditional distortion metric (STDM)

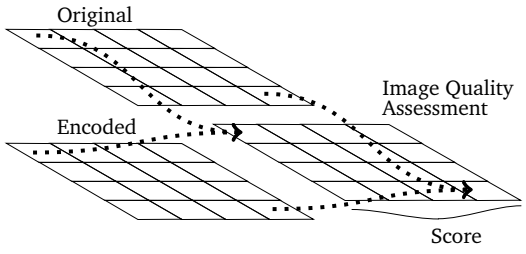


Figure 4.2 Image quality assessment (IQA)

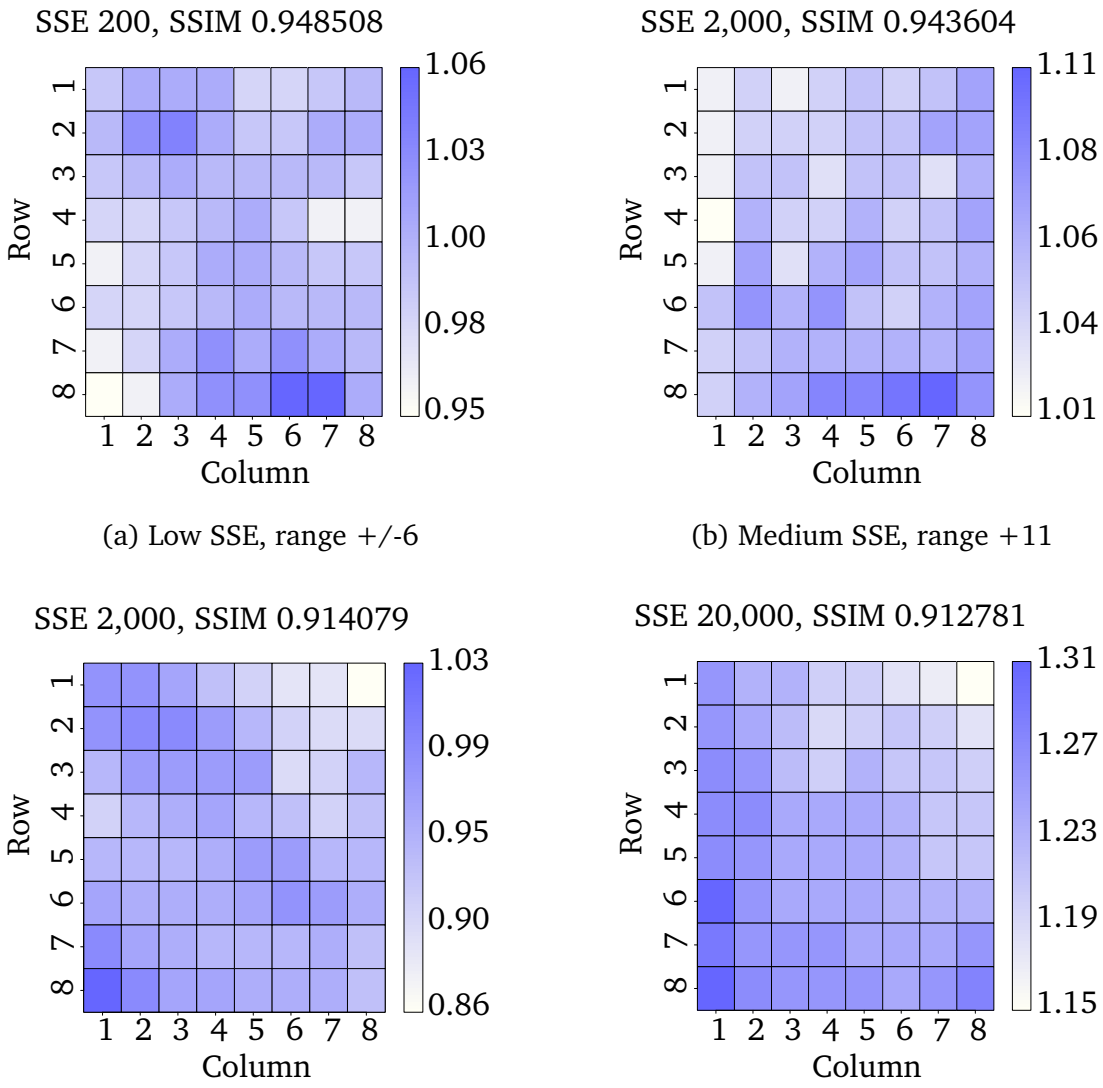
and Figure 4.3b and similarly at Figure 4.3c and Figure 4.3d, these pairs show that for similar SSIM the SSE can differ. Conversely, for the same SSE, the SSIM can differ when one considers Figure 4.3b and Figure 4.3c. These differences in scores are related to the statistical methods of SSIM. Figure 4.3 illustrates that the spread of relative differences are a contributing factor. Unfortunately, these related findings are too limited to make a specific conclusion and so more observations are required. Therefore, by conducting experiments at the sub-block level, it is possible to gather additional observations from a variety of video sequence sources. Then the analysis of these observations can provide a basis from which SSIM can satisfy the triangle equality rule ($\triangleq$).

4.1.3 Triangle equality rule: STDM vs. SSIM

Existing PVC solutions that use SSIM to rescale quantisation need to overcome the compatibility issue of the respective distortion measures. From a mathematical perspective, the problem is centred around satisfying the $\triangleq$ when resolving SSIM to the existing distortion assessment. This issue can be highlighted with an example shown in Table 4.1, where a set of 8x8 array of pixels are shown for original, prediction and the differences. Consider these as a set of image triplets, original (A), predicted (B) and differences (C). The $\triangleq$ stipulates that given these three sides (A,B and C), no two side should be greater than the third, which means that through the various combination of these image triplets they should be compatible once assessed. In this case, the pairs are $`A,B'$, $`B,C'$ and $`A,C'$. When these are assessed using SATD for non-perceptual and SSIM for perceptual, the results shown within the summary table illustrate that for SSIM the score is very different, $>$6.7k%, demonstrating that SSIM does not satisfy the $\triangleq$.

4.1.4 Existing non-perceptual distortion metrics

As illustrated, an IQA like SSIM is unable to support the $\triangleq$. This research will consider mapping SSIM within an existing distortion metric space which does fulfil the $\triangleq$, principally similar to an existing approach (Bhat, Richardson and Kannangara, 2010). The reasoning behind this method of replacing the STDM like SAD and SSE with an IQA is because STDMs are dimensionless. STDMs operate as a series of calculations that do not factor the surrounding pixels. STDMs treat a given



(a) Low SSE, range +/-6

(b) Medium SSE, range +11

(c) Medium SSE, range +3 to -14

(d) High SSE, range +15 to +31

Figure 4.3 Arrays of 8x8 sub-blocks with conflicting SSIM and SSE scores

pixel pair (source and reconstructed) uniformly, this allows them to be tractable and allow the distortion assessment to be classed as a metric. The independent pixel level evaluation by STDMs makes it scalable and unaffected by adjacent pixel differences. In turn, STDMs are unable to appreciate the significance of inherent perceptual visual clues like structure or texture within the sub-blocks. In computational terms, both SAD, SATD and SSE depend on for-loops, with SAD utilising an absolute function, SATD, which utilises the Hadamard transform and SSE requiring a squaring operation.

$A_{Original}$ - Original 8x8 Block

100	106	110	111	114	118	119	118
97	98	104	110	114	117	119	119
108	105	103	107	112	116	118	120
123	118	113	111	111	114	116	118
126	125	120	119	117	114	115	117
126	126	123	121	121	119	117	117
125	126	125	124	123	122	119	118
120	123	125	126	125	124	121	119

$B_{Prediction}$ - Prediction 8x8 Block

116	116	116	116	116	116	116	116
116	116	116	116	116	116	116	116
116	116	116	116	116	116	116	116
116	116	116	116	116	116	116	116
116	116	116	116	116	116	116	116
116	116	116	116	116	116	116	116
116	116	116	116	116	116	116	116
116	116	116	116	116	116	116	116

$C_{Difference}$ - Difference 8x8 Block

-16	-10	-6	-5	-2	2	3	2
-19	-18	-12	-6	-2	1	3	3
-8	-11	-13	-9	-4	0	2	4
7	2	-3	-5	-5	-2	0	2
10	9	4	3	1	-2	-1	1
10	10	7	5	5	3	1	1
9	10	9	8	7	6	3	2
4	7	9	10	9	8	5	3

Summary of SATD and SSIM scores for pairs of 8x8 blocks

	$`A,B'$ (1)	$`B,C'$ (2)	(1)+(2)	$`A,C'$ (3)	Diff	Diff (%)
SATD	509	1859.5	2368.5	2339.5	29	1%
SSIM	0.532	0.014	0.546	0.008	0.538	6725%

Table 4.1 Example highlighting SSIM's triangle equality issue, using 8x8 blocks

4.1.5 Technical challenge of using SSIM

While STDMs need only differences to calculate a score, implementing SSIM at the sub-block level means that the original and reconstructed pixel values are required. This means that at this point where the differences would normally be calculated, the original values should be passed alongside the differences. Within the JM H.264/AVC codebase, the differences are calculated within each STDM call, however there are specific places where the differences are pre-calculated and passed to the STDM. During mode decision, where RDO takes place, the encoder assumes the STDM is a true metric, as these calls occur when the encoder decides if a single large 8x8 block is suitable in place of four smaller 4x4 blocks. During this RDO process, having SSIM derived scores which are STDM compatible across block sizes can give rise to PVC at the native sub-block level. Thus, when the encoder chooses a larger block size upon it has similar perceptual loss than four 4x4, which needs fewer bits for storing the signalling and residual information. This highlights a crucial aspect of why the $\triangleq$ must be adhered to. If the scaling of SSIM within the SATD distortion metric space is not done correctly, this can favour a larger block size over several smaller blocks, making it is crucial to observe how SSIM operates at the sub-block level.

4.2 Design

As the introduction stated, this chapter consists of exploring the SSIM and STDMs relationship at the sub-block level. As such, this has been marked as three experiments of: observation gathering, identifying component(s) of SSIM-STDM relatiopship and producing a low complexity model to scale SSIM scores to STDM compatible value. The reason for this is to provide a means to ensure SSIM scores are STDM compatible, which would allow perceptual assessment at the sub-block level. However, as explained in 'Section 4.1 Related findings', the wide dissimilarity between pairs of SSIM and STDMs scores for the same pixel arrays indicate that something is not fully understood. There are two conflicting theories, a simplistic non-linear response, which is used in existing SSIM based PVC solutions, and another theory, of a geodesic triangle where different paths exist (Brunet et al., 2012; Horé and Ziou, 2013). More importantly, this relationship must be understood so to enable a low complexity model to scale SSIM to STDMs, which will allow perceptual assessment at the native sub-block level. Therefore, each of the three experiments will contribute towards improving the understanding between SSIM and STDM. The first experiment will seek to resolve this conflicting theories by exploring the SSIM and STDM relationship at the sub-block level, gathering observations which can demonstrate which of the theories is applicable. In the second experiment this SSIM-STDM relationship must be understood beyond the universal bounded region (UBR) covered by the gathering of observations. Finally, in the third experiment, the UBR understanding will be used to produce a low complexity model, in which to scale SSIM scores to be STDM compatible. In all, this will allow a novel PVC solution where SSIM is applied at the sub-block level, which should enable HVS friendly encodings.

4.2.1 First experiment: observing SSIM

Existing SSIM based PVC solutions work outside the native sub-block level, because of the high complexity in making SSIM scores compatible. Under existing STDM distortion assessments, operating with the pixel array of differences is sufficient to calculate distortion metric scores. This means that any investigation into IQA at the sub-block level requires adapting the encoder to support perceptual

calculation at the same point as STDMs. In particular, undertaking this observational study at the prediction stage allows gathering observations for all sub-block candidates which would be used to evaluate the same original sub-block. This will give greater means of assessment based upon the same source information.

The prediction stage assesses several candidates per sub-block both for intra and inter coding. As these prediction candidates include intra modes and inter motion vectors, this increases the volume and dynamic range of scores. This allows for the possibility to map pairs of responses for each candidate by their existing STDM and IQA score. Mapping perceptual scores to STDM can illustrate the $\triangleq$ issue faced by SSIM, yet with sufficient observations it should be possible to model the response curve of SSIM to compatible STDM scores. This means, gathering these observations will be significant in providing a basis from which subsequent experiments can be conducted towards modelling this response.

Relevant literature for conducting observational study

Observational studies into IQA have largely been performed as a means to optimise PVC solution, operating at the sequence, frame or block level. These studies have established a non-linear curve relationship between SSIM and STDM ('SSIM-STDM') by way of paired assessment results (Bhat, Richardson and Kannangara, 2010; Horé and Ziou, 2013; G.-L. Wu et al., 2013). This approach can be seem as simplistic when one considers the related findings in Figure 4.3. This means that the current evidence is limited and so more observations from a variety of sources are needed. Importantly, existing literature does not explore the sub-block level, where higher volume of observations can be gathered than the methods used previously.

Hypothesis for observational study

Following from the findings of the literature cited above, it would be fair to state that a non-linear response is expected. This can be justified further by the understanding that perceptual is modelled around the HVS, which has a non-linear sensitivity to frequency. Conversely, STDMs are designed for uniform differences, irrespective of content. This means candidates with the same STDM score may have different perceptual scores depending upon the content, as shown in figure 4.3. As a consequence, the 'SSIM-STDM' relationship is expected to be a region of shared

space than a non-linear line. This will reinforce the description of the geodesic-$\triangleq$ description of the relationship between SSIM with STDMs by investigating SSIM at the sub-block level, which includes 8x8 and 4x4 pixel array size. SSIM is designed to run optimally around 8x8, while a 4x4 window size is more susceptible to a single dominant pixel, meaning it is expected that 8x8 and 4x4 'SSIM-STDM' will have different responses (Brooks, X. Zhao and T. Pappas, 2008). The SSIM and STDM relationship is expected to be a region of space where observations may complement or conflict, where, SSIM is constrained by the index range of ± 1, while STDMs are limited by the spatial size and bit-depth. This means that the shared region of space, can be described as the universal bounded region (UBR). It is important to be able to understand and model the UBR, that way it can enable an accurate representation of the SSIM-STDM relationship.

4.2.2 Second experiment: understanding the universal bounded region (UBR)

The first experiment introduced a new concept, the universal bounded region (UBR), defining SSIM-STDM relationship as a shared region of distortion space. Being able to understanding the UBR can potentially resolve the $\triangleq$ issue of SSIM. This has the potential to allow perceptual distortion assessment to occur at the native sub-block level. In particular, it would allow SSIM scaled to a STDM compatible distortion cost, similar to rate-control with quantisation. Therefore, this experiment is to discover the underlying component(s) of the UBR, which will contribute towards mapping SSIM scores to STDM within shared space of the UBR.

Potential for manipulating the UBR

The UBR represents a suitable shared space to scale SSIM as a compatible STDM score. Understanding what leads to the UBR being shaped as is, can allow the UBR to be manipulated based upon the content, bandwidth or application. This means identifying the component(s) by which to manipulate the scaling of SSIM. In turn, this can be represented within the classic R-D equation, similar to how λ is used to regulate the level of quantisation, the distortion cost could be manipulated. To regulate this response of distortion based upon the UBR, the symbol κ can be used to represent the perceptual scale on the distortion score as shown in Equation (4.1),

$$J_{min\ energy} = \lambda_{quant} \cdot R_{bit\ rate} + \kappa \cdot D_{dist\ metric} \tag{4.1}$$

This means that κ can complement the operation of λ with potentially finer control adapted per candidate.

Identifying factors for scaling SSIM

While the first experiment is designed to demonstrate the shared space of the UBR, it does not explain what causes this behaviour. In order to understand what may be influencing the UBR, analysis must occur further than the sub-block level, at the pixel level to identify factors for scaling SSIM. By extracting the pixel luma values of sub-blocks, this allows breaking the calculation of SSIM and SATD scores in to stages. An example of this is shown in Table 4.2 which summaries the raw luma values and associated scores. The table illustrates that for similar SSIM there is a change in covariance, which suggests a wider experiment focused on capturing covariance is required.

Valid range of covariance values

As shown in Table 4.2, small changes in SSIM and SATD may result in a wide range of covariance. Therefore, before applying covariance into a model of the UBR, the potential range should be understood and compared to that observed in the first experiment. This can be explained by understanding the definition of covariance. Covariance reflects the degree of similarity, as a normalised value of the sum product of each sample from their respective mean as shown in Equation (4.2).

$$\sigma_{OrigRec} = \frac{\sum_{i=1}^{N} (Orig_i - \mu_{Orig}) \times (Rec_i - \mu_{Rec})}{N} \tag{4.2}$$

8x8 Pairs	Pair 1	Pair 2	Pair 3
Mean Diff	4.25	1.98	0.52
Mean Diff (%)	2%	1%	0%
Covariance	3139.77	1.55	0.17
SSIM	0.97	0.92	0.93
SATD	415.50	231.50	176.50

Table 4.2 Covariance and SATD relationship - Calculated Manually using Raw Pixel luma 8x8 Blocks Pairs

where $\sigma_{OrigRec}$ represent covariance (also written as $Cov(x,y)$), $Orig$ and Rec are the pixel arrays of the original and reconstructed blocks respectively, and N is the number of pixels in the block. Consequently, covariance can be seen as a measure of the texture activity between original and reconstructed pixel arrays.

From Equation (4.2), it is possible to determine the range that covariance will operate within. By assuming an extreme case, where pixels are oscillating between theoretical high and lows, the mean pixel value across the block will be exactly halfway. In terms of 8-bit grey-scale, luma pixel value, this will result in a low of zero, a high of 255, and average of 127.5. Taking 127.5 away from either extremes of zero and 255 leaves with $\pm 127.5^2$, and since the pattern of oscillation is repetitive, both the summation and denominator cancel each other out. This makes the covariance score range $\pm$16,256.25 for an 8 bit luma, which is very large, however, the example taken was very extreme and the likelihood that occurring is very low. A more suitable range would be half of the theoretical maximum, $\pm 8k$, from the nature of how covariance operates this would probably mean that high pixel values are incorrectly represented. Therefore, modelling SSIM-STDM relationship by covariance will ensure that SSIM scores which represent bright regions are more likely to have a higher STDM score. This is because covariance scales non-linearly, which means that differences represent more for brighter pixels.

Predefining SSIM's bit-depth related constants

The sub-block level is computationally sensitive due to the volume of calls, meaning that SSIM must be made efficient. Previous attempts at making SSIM efficient have been based upon providing making unfair assumptions (Rouse and Hemami, 2008). One means to minimise the computational overhead of SSIM, is to have its bit-depth related constants C_1 and C_2 predefined at the initialisation of the video encoder, than for every call to SSIM. In Equation (4.3) SSIM is shown as

$$SSIM(Orig, Rec) = \frac{(2\mu_{Orig}\mu_{Rec} + C_1) \cdot (2\sigma_{Orig,Rec} + C_2)}{(\mu_{Orig}^2 + \mu_{Rec}^2 + C_1) \cdot (\sigma_{Orig}^2 + \sigma_{Rec}^2 + C_2)} \tag{4.3}$$

where the constants, C_1 and C_2, should smooth out variation, however the constants are derived upon the level of bit depth, luma, typically 8 bits and

additional constants called K_1 and K_2 that are stated as $<< 1$, (where $<<$ stipulates far smaller than). However, coupled with the implementation of SSIM in JM MPEG4/AVC (Sühring, n.d.), C_1 and C_2 can be defined as shown in equations 4.5 and 4.6, assuming 8 bit luma (greyscale bit depth).

$$\text{Let BitDepth} = 8,\ K_1 = 0.01 \,\&\, K_2 = 0.03 \tag{4.4}$$

$$\Rightarrow C_1 = (0.01 * 255)^2 = 6.5025 \tag{4.5}$$

$$\Rightarrow C_2 = (0.03 * 255)^2 = 58.5225 \tag{4.6}$$

The vast majority of encoded video has 8 bit luma (grey-scale) information, and it is accepted that the HVS is more sensitive to grey-scale than colour information (H. Wu and Rao, 2005). This means these derived values of C_1 and C_2 from equations 4.5 and 4.6 respectively, can be described as constants, or set-up upon initialisation of the encoder. Overall, this is a minor change, however, due to the iterative nature of assessment in video coding, these should account towards complexity saving.

4.2.3 Third experiment: design a low complexity scaling of SSIM based upon the UBR

The third experiment is about applying the understanding of covariance towards modelling the SSIM-STDM relationship. For existing PVC solutions, models for the SSIM-STDM relationship described it as a single non-linear curve leading to high complexity forms of scaling (Bhat, Richardson and Kannangara, 2010; Y.-H. Huang, Ou, Su et al., 2010). By identifying covariance as a variable to scale SSIM, and that the UBR is a 2D space, this gives more flexibility in creating a new means of scaling SSIM scores. Also, since covariance is calculated as part of SSIM, this allows component re-use, SSIM can be scaled using its own statistical property of covariance. This allows the possibility for a low complexity scaling for SSIM to occur for distortion assessments at the sub-block level. Therefore, this experiment will investigate how to model the UBR as a low complexity design to scale SSIM as part of a proof-of-concept design.

Classifying the UBR into zones

The third experiment uses covariance to model the UBR such that SSIM scores are scaled to be SATD compatible while using low complexity techniques. The UBR covers a large range of covariance, to manage this the UBR model was set with a limited covariance range of 0 to 8000 and furthermore covariance was broken into six zones. This limited covariance range corresponds to where 1-SSIM is between 0 and 1, while for the remaining range of 1 to 2 the encoder would revert back to SATD. This design choice was done to avoid the complexity of negative covariance when 1-SSIM $>$ 1 and to keep this model as a proof-in-concept experiment. The six covariance zones to cover the UBR were split in five places where covariance was 150, 300, 600, 1350, 3014 respectively. Apart from the choice of 3014 which was set to avoid confusion with 300, the shape of the observations lent itself to the respective covariance band. When these covariance bands were adapted to in terms of log_{8k}, they represented 0.56, 0.63, 0.71, 0.80 and 0.89. Figure 4.4 gives an indication of how the different covariance zones were formed, where zones with lower values of covariance covered a broader range of 1-SSIM values. In Figure 4.4, the zones correspond one or many colours as they are shown using log_{8k} scale. In order to model these zones using low complexity techniques, the respective trends for these zones were traced, then scaled based upon linear graphs and processor friendly values. This meant that SSIM and covariance scores would undergo a different set of linear equations. Overall, this produced the proposed implemented framework using psuedo-SSIM called local hybrid pseudo-SSIM-SATD (LHPSS), as shown in Figure 4.5.

Overview of proposed framework: local hybrid pseudo-SSIM-SATD (LHPSS)

The pseudo-SSIM model consisting of linear operations and presented as part of the proposed frameworkd LHPSS is shown in Figure 4.6. The scaling of SSIM to pseudo-SSIM is designed with adds and shifts where possible, with values chosen to be binary friendly, thus promoting processor friendly techniques. For reasons of simplicity during modelling SSIM is first converted to (1−SSIM)×1000, then multiplied by covariance to form the covariance scaling cost. Inside the mean difference cost for 8x8 blocks, a multiplication operation is also used to provide a degree of additional variation. The LHPSS is a partial model of the

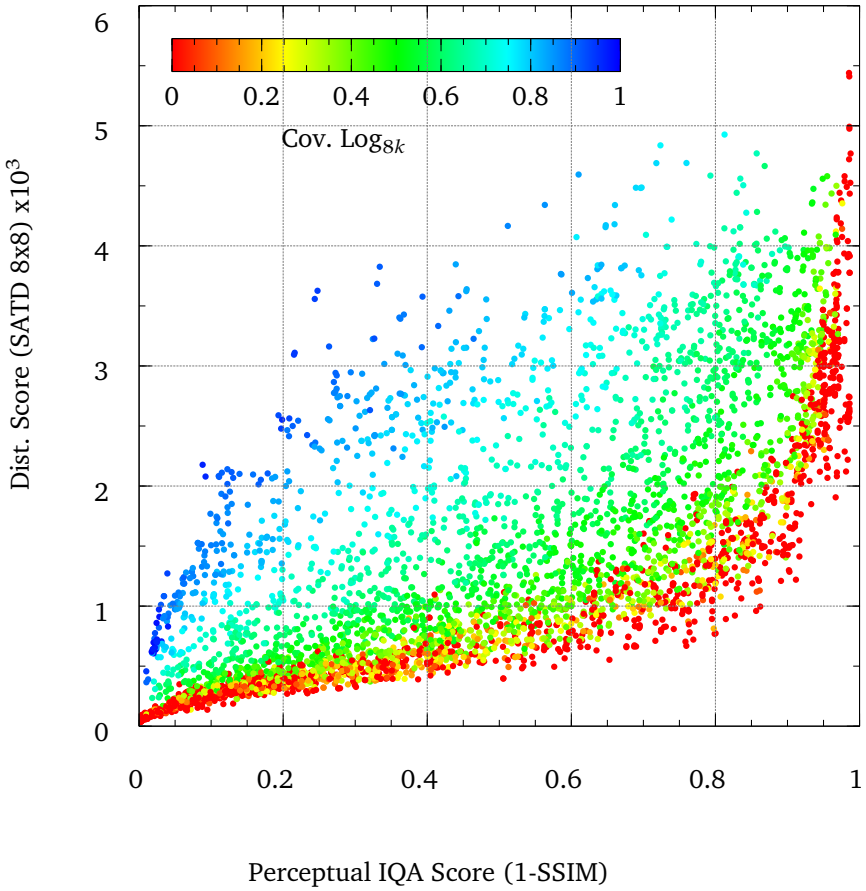


Figure 4.4 SSIM vs. SATD (8x8) with Covariance used for modelling UBR by way of zones and line graphs with scaling coefficients.

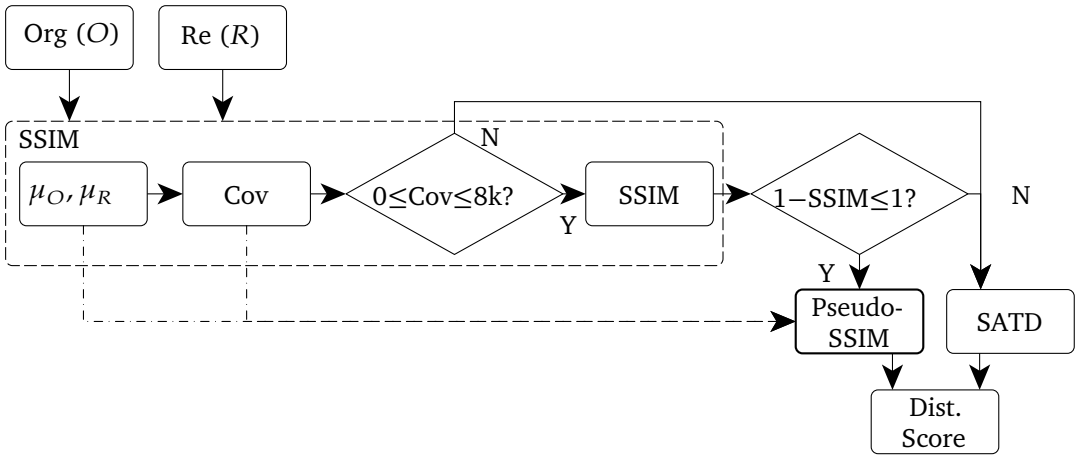


Figure 4.5 Flowchart of local hybrid pseudo-SSIM-SATD distortion metric.

UBR called pseudo-SSIM, where both covariance and difference of means are used in a low complexity design. Pseudo-SSIM gathers pre-defined weightings for the respective covariance zones from a look up tables (LUT). The use of ΔCov is relative covariance with the respective covariance zones and used extensively, especially with weightings, mean difference and SSIM. Compared to existing methods of scaling SSIM scores which have logarithmic operations, pseudo-SSIM operational block diagram illustrates a low complexity scaling. In all, pseudo-SSIM scales SSIM by using three divides, four multiplies, four additions and two subtractions (including mean difference and ΔCov).

A detailed look at UBR model: pseudo-SSIM

The internal processes involved to calculate pseudo-SSIM are shown in Figure 4.6. They consist of three main parts, the mean difference cost, the covariance scaling cost and the SSIM scaling cost and together they sum up to output a pseudo-SSIM score. Of the three, covariance scaling is the most complex, however, it is at the earlier stage when covariance is obtained that illustrates how the operation is driven by the covariance value.

In order to simplify the overall process, covariance based zones have been constructed with their own set of SSIM profiles. There are six major zones, when covariance is passed it takes the form of ΔCov, the zones base value, i.e. for the second zone of 150 to 300, 150 has been subtracted and thus the relative covariance value is made available. The purpose of such a technique is that subsequent range within the particular zone, between the base value and the ΔCov can then be factored and added as part of the covariance scaling cost.

With pseudo-SSIM, SSIM must be in integer form, for reasons of computing efficiency, therefore, at the expense of losing some accuracy, multiplied by one thousand before being used in covariance scaling cost and SSIM scaling cost. Also, depending upon block size, the mean difference cost is calculated differently. In terms of the 8x8 model, the mean difference cost is added to the final value. While in 4x4, covariance scaling can play more significant role and therefore, there is a secondary ΔCov factor that is added to the final result.

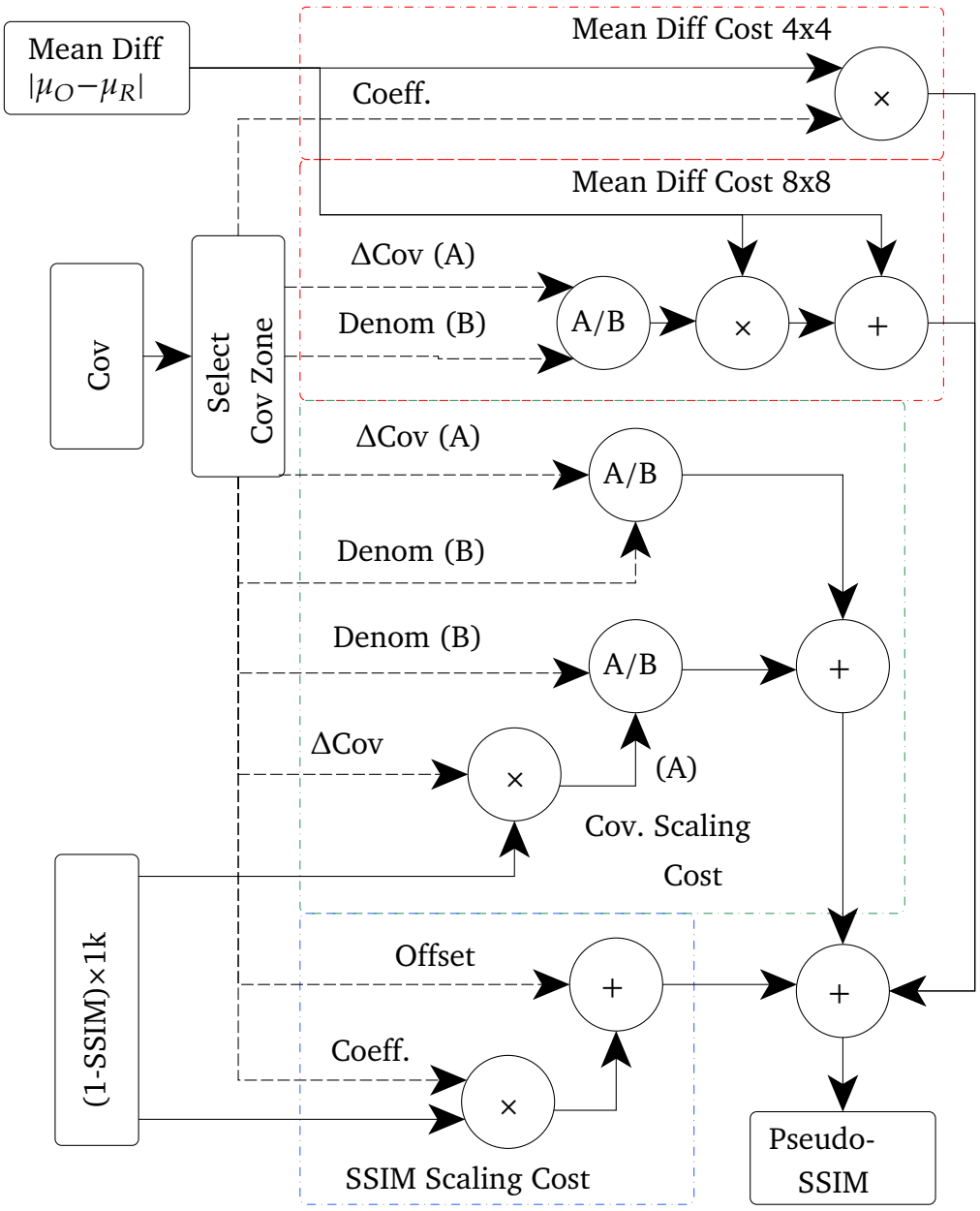


Figure 4.6 Operational block diagram of pseudo-SSIM

4.3 Methodology for testing experiments

To conduct these experiments JM H.264/AVC (version 18.4) was modified (Sühring, n.d.). During the first two experiments where observations were gathered, the codebase would be adapted to store prediction candidates scores, using both

STDM and SSIM. This meant that the proposed UBR can be tested and later prove whether the UBR could be steered by covariance. For the third experiment, the codebase was modified such that the model of the UBR based on covariance is applied on SSIM scores to provide STDM compatible scores. The video sequences used in the first experiment was Quarter Common Intermediate Format (QCIF) and CIF resolution. For the second experiment CIF video resolution was used, and finally for testing the model HD resolution was used. In all, the method centred around modifying the codebase, to gather data based upon video sequences than to simulate. These experiment gathered observations during prediction, were a high volume and broad variety of results occur. This is important, as that way it ensured that the UBR model produced in the third experiment can be said to be applicable to a variety of situations.

4.3.1 First experiment: method to capture observations

The methods used in this first experiment revolved around assessing the same candidate with three STDMs and SSIM. In this case, SATD 8x8 and SATD 4x4 were modified to also assess the same candidate with SSIM, SAD and SSE forms of assessment. SATD was the preferred distortion metric as it offered the most significant complexity whilst being native to the prediction encoder workflow. Initially, tests were orientated around the coding types of Intra and Inter, then subsequently, they centred upon video sequences using SATD as the sole distortion metric.

Set-up for observational study

This experiment was conducted on JM H.264/AVC version 18.4, which supports sub-block sizes of 4x4 and 8x8 (Sühring, n.d.). JM H.264/AVC was modified to support passing original and reconstructed pixel arrays to the final distortion metric stage so that SSIM could be calculated. Hadamard 8x8 and 4x4 function in JM H.264/AVC were modified to support the additional assessments (SSIM, SAD and SSE). These assessments were calculated individually and then saved to a file containing scores from each distortion assessment. The calculation of SSIM was taken from the implementation found in the JM. H.264 decoder. In all, each time the Hadamard 4x4 or 8x8 function was called the modified encoder appended the respective distortion scores to the log file. Due to the volume of results only the

first three frames were encoded using main profile random access configuration, which allowed intra, inter predictive and bi-predictive frames to be encoded. With QCIF video resolution, 900k and 700k observations were captured for 4x4 and 8x8 respectively during inter frames. Later with CIF video resolution these values raised to over 4 million, which produced its own technical challenges.

Test configurations

This experiment was designed to observe both multiple STDM-SSIM assessments for a single video sequence and a single STDM-SSIM for mulitple video sequences. This meant that in the initial experiment 12 set of tests were conducted, using 3 STDMs (SSE, SAD, SATD) × 2 coding types (intra or inter) × 2 block size (4x4 and 8x8), where the single video sequence was used Foreman, a quarter common interchange format (QCIF) resolution video. While for the later experiment, SATD and inter coding observations was selected, which allowed multiple video sequences for the same number of tests. The decision to use inter coded frames was because they represent a more dynamic range than intra, and inter reflects the majority of coding applications of low delay P and random access. A variety of video sequences were selected, six in total, all of CIF resolution; Bus, Flower, News, Stefan, Tempete and Waterfall. Overall, the number of results for this second part of the experiment were 12, 6 video sequences × 2 block sizes.

Data processing of observed pairs

The encoder recorded all SATD 4x4 and 8x8 scores to a file, along with the corresponding STDM and SSIM scores. Despite limiting the video sequence to the first three frames, the nature of inter coding means a high volume of observations were gathered. The resulting data had to be processed using R, an open-source statistical tool, which allows large number of observations to be managed (R Core Team, 2014). The data was process to contain unique pairs and stored as a comma separated values (CSV) file. This CSV file was imported to a scientific plotting tool called Veusz, from which graphs used in the results were created (Sanders, 2015).

4.3.2 Second experiment: identifying components of the UBR

In this second experiment, the same methodology as the first experiment was applied, where pairs of results SATD and SSIM were logged, however, with additional information, representing the component of the SSIM equation. These

additional values included mean, variance and covariance for the respective sub-blocks. Subsequently, these components of SSIM were analysed against the SATD score using the statistical package of R to identify any correlation. Finally, to compared the significance of covariance as a component of the UBR it was compared against a perceptual model. The JND was chosen as it represents pixel domain and is aimed at image coding, where the JND equation in Equation (4.7) is as follows:

$$JND(x,y) = \begin{cases} 17 \times (1 - (\frac{bg(x,y)}{128})^{\frac{1}{2}}) + 3 & bg(x,y) < 127 \\ \frac{3}{128} \times (bg(x,y) - 127) + 3 & bg(x,y) \geq 127 \end{cases} \tag{4.7}$$

where $bg(x,y)$ is the background luminance, in this case the higher of two pixel pair values. This allowed evaluating the UBR component in terms of a perceptual perspective.

4.3.3 Third experiment: modelling the UBR

Using the same JM H.264/AVC code-base, the proposed pseudo-SSIM model and the framework of LHPSS was implemented. Under H.264/AVC both 4x4 and 8x8 sub-blocks for SATD were modified to support LHPSS. This design provides a means to utilise SSIM in a less constrained manner with pseudo-SSIM promoting reuse of SSIM components in terms of scaling and using linear equations to ensure processor friendly operations. Overall, this allows pseudo-SSIM to be seen as a general solution applicable to any block-based codec, including the HEVC (Sullivan et al., 2012). Compared to existing PVC solutions which operates on key frames only (one frame per GOP), or at RDO mode decision (block level) this provides an individual candidate assessment.

4.4 Results

These results in this section illustrate the SSIM and STDM relationship across different coding types, block sizes and video sequences. Also, the UBR is shown by way of covariance and how it compares to a video frame compared to perceptual model of JND. Finally, the implemented UBR scaling model is shown with visual and numerical results, showing where changes are occurring and to what amount respectively.

4.4.1 Results gathered from observational study

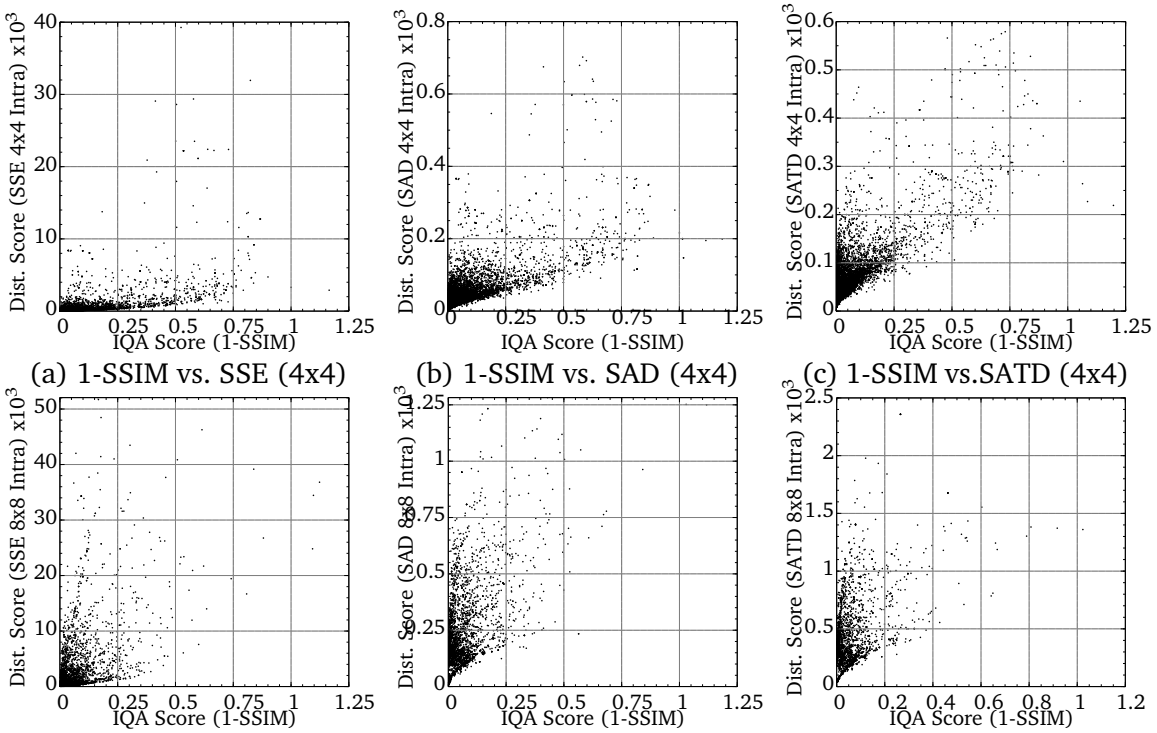
Overall, the results are presented by coding type and by video sequence, as shown in Figures 4.7 to 4.9 for 4x4 and 8x8 sub-blocks. The results are presented as a series of scatter graphs of perceptual IQA (1-SSIM) versus the respective STDM. IQA of SSIM is presented as one minus SSIM (1-SSIM) in order to have the origin represent zero distortion for both perceptual and STDMs. Likewise, the maximum scores on the axis's relate to the highest amount of distortion for the respective form of assessment. Finally, Figure 4.9 illustrates the results by individual video sequences. For Figures 4.7 and 4.8 each column of graphs represent different distortion metric by which the same observation was assessed by, SSE, SAD and SATD respectively.

4.4.2 Results for identifying components of the UBR

The resulting graph Figure 4.10 illustrates the effect covariance has on the scores of the UBR. Each observation was highlighted by its covariance value, which was first shifted (+580) to be all positive then normalised to (Log_{16k}) within the maximum positive range. As a comparison, the covariance values were used against a reconstructed frame and JND on the original and this is shown in Figure 4.11 and Figure 4.12 respectively.

4.4.3 Results for pseudoSSIM

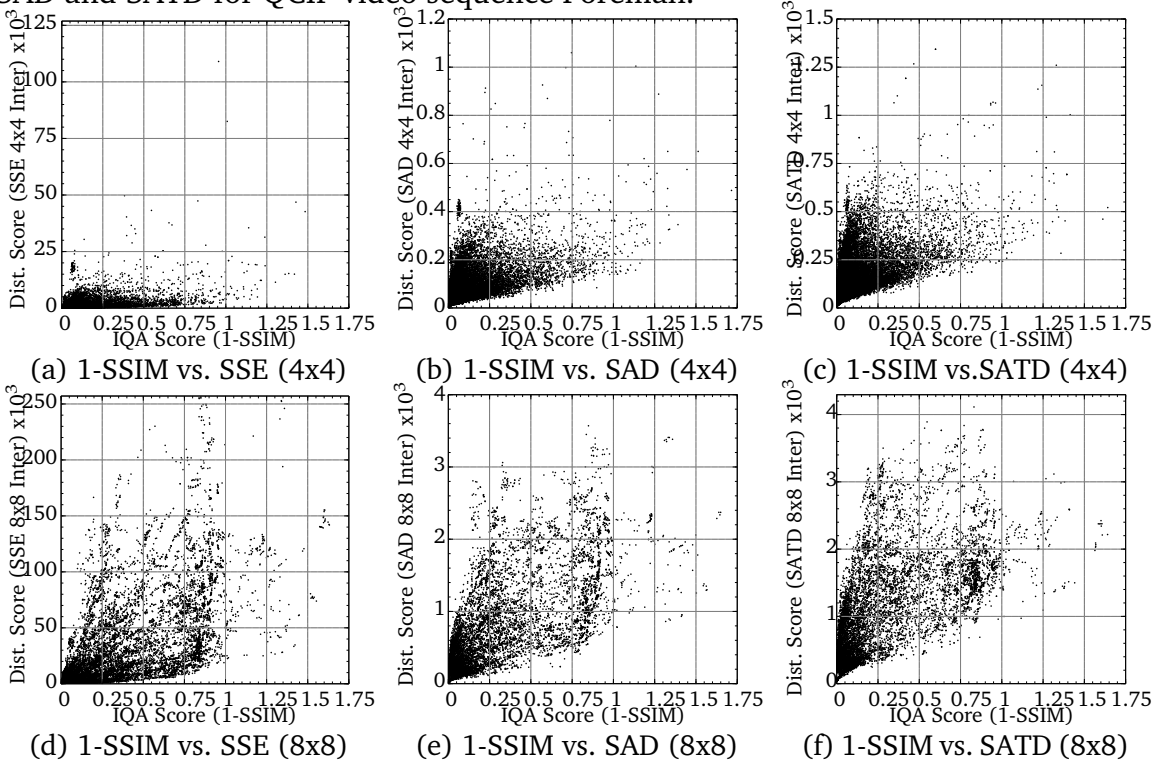
The LHPSS was implemented on the JM H.264/AVC codebase version 18.4 on the SATD distortion metric, called during intra and inter coding. Tests were performed using a system with an Intel Core i7 CPU 920 processor operating at 2.67GHz and 7GB of RAM. The results are presented in two forms, the first is a graphical representation of the initial intra frame for of two HD (1080p) video sequences, CrowdRun and Sunflower as shown in Figure 4.13. While the second form is tabular, in the form of Table 4.3, where both video sequences are encoded under random access and low delay P coding structures. The graphical form of results the block sizes are highlighted as either red, green or blue to represent 4x4, 8x8 or 16x16 blocks respectively. As the initial frame for both coding structures are the same, intra, Figure 4.13 is applicable to both random access and low delay p. The tabular set of results are shown with four different quantisation parameters



(a) 1-SSIM vs. SSE (4x4) (b) 1-SSIM vs. SAD (4x4) (c) 1-SSIM vs.SATD (4x4)

(d) 1-SSIM vs. SSE (8x8) (e) 1-SSIM vs. SAD (8x8) (f) 1-SSIM vs. SATD (8x8)

Figure 4.7 IQA vs. STDMs from 4x4 and 8x8 intra blocks. SSIM, plotted against SSE, SAD and SATD for QCIF video sequence Foreman.



(a) 1-SSIM vs. SSE (4x4) (b) 1-SSIM vs. SAD (4x4) (c) 1-SSIM vs.SATD (4x4)

(d) 1-SSIM vs. SSE (8x8) (e) 1-SSIM vs. SAD (8x8) (f) 1-SSIM vs. SATD (8x8)

Figure 4.8 IQA vs. STDMs from 4x4 and 8x8 inter blocks. SSIM, plotted against SSE, SAD and SATD for QCIF video sequence Foreman.

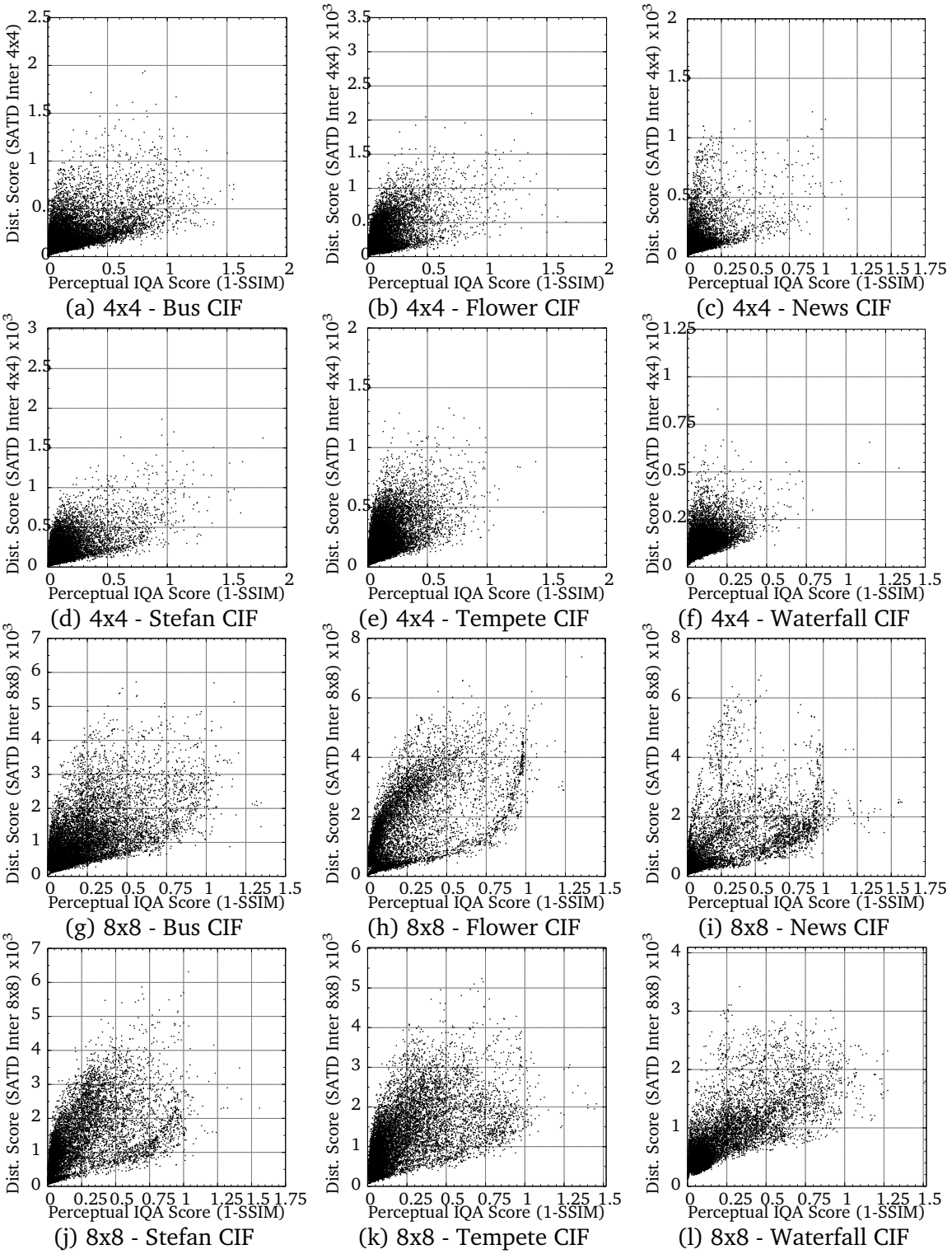


Figure 4.9 IQA vs. STDM from 4x4 and 8x8 Inter Blocks. SSIM, plotted against SATD for CIF video sequences (Bus, Flower, News, Stefan, Tempete and Waterfall).

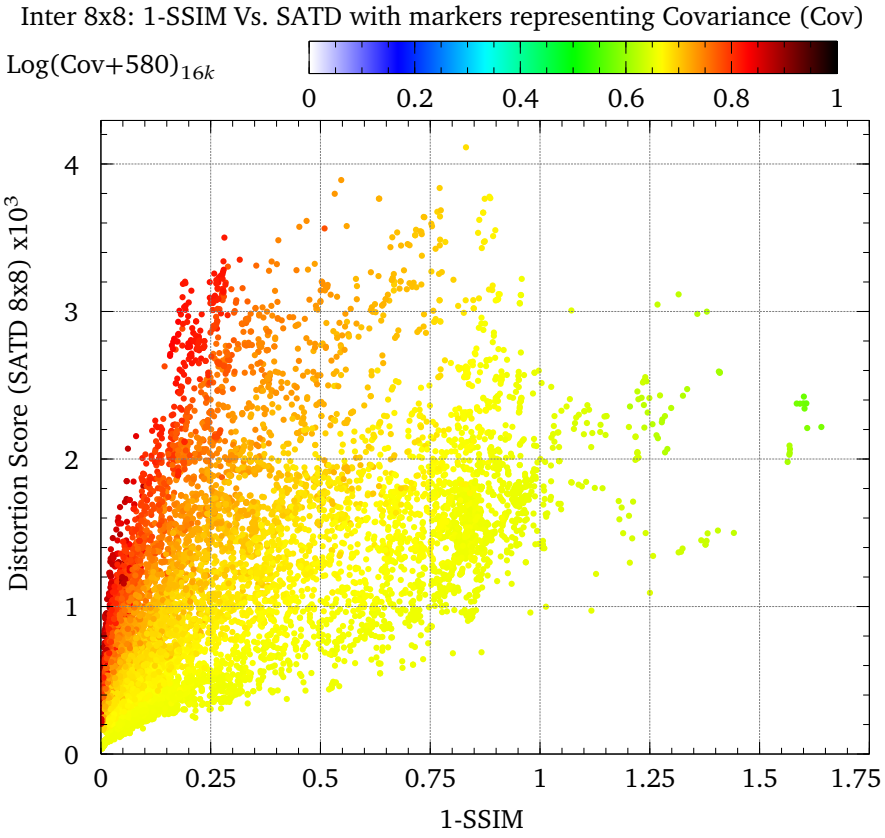


Figure 4.10 SSIM vs. SATD (8x8) with Covariance.

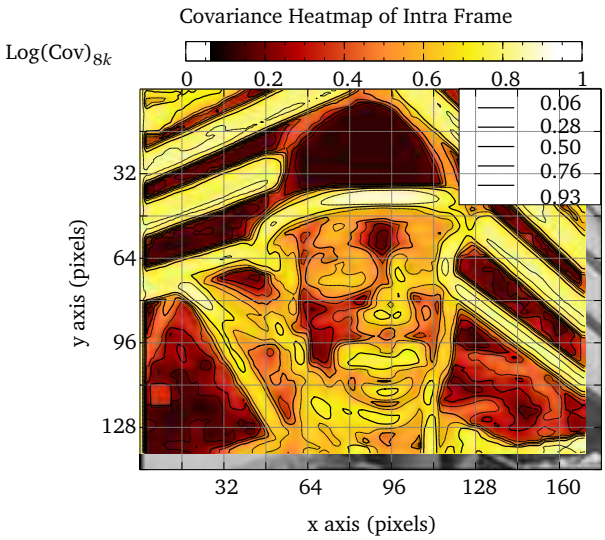


Figure 4.11 Covariance Heatmap of Intra Frame (Foreman frame 0 QCIF).

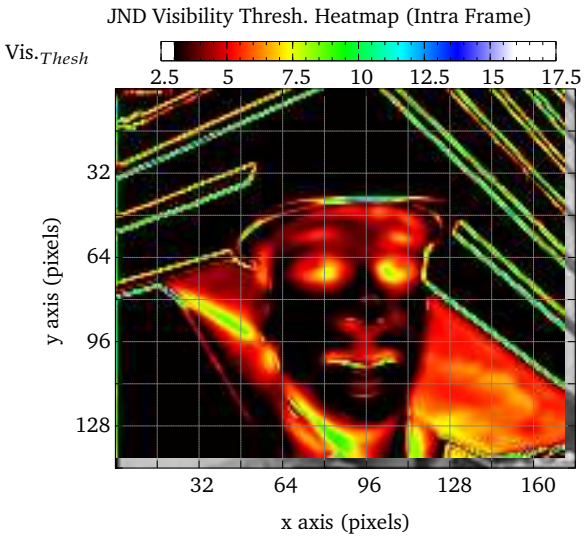


Figure 4.12 Just Noticeable Distortion (JND) Visibility Threshold of Intra Frame.

(QP) values set, QP22, QP27, QP32 and QP37, along with the resulting total time, luma-PSNR, luma-SSIM and total bit usage.

Random access (coding structure: IbBbBbBbP)

CrowdRun	QP22	QP27	QP32	QP37	Ave.
Total Time	19.70%	19.70%	12.67%	21.07%	18.29%
Luma-PSNR	-0.35%	-0.55%	-0.65%	-0.57%	-0.53%
Luma-SSIM	-0.06%	-0.20%	-0.44%	-0.68%	-0.34%
Total Bits	1.95%	1.38%	0.53%	-0.54%	0.83%

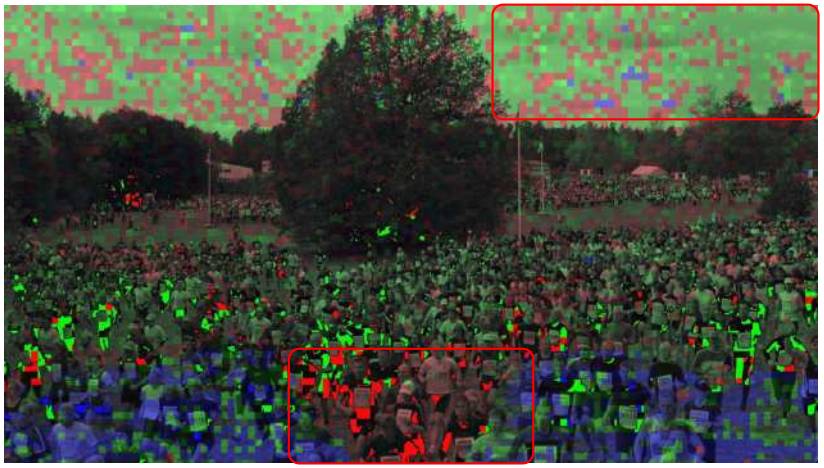
Sunflower	QP22	QP27	QP32	QP37	Ave.
Total Time	21.16%	21.91%	23.50%	24.81%	22.85%
Luma-PSNR	-0.19%	-0.30%	-0.34%	-0.43%	-0.31%
Luma-SSIM	-0.03%	-0.07%	-0.16%	-0.39%	-0.16%
Total Bits	0.57%	-1.99%	-5.45%	-9.49%	-4.09%

Low delay p (coding structure: IPPP)

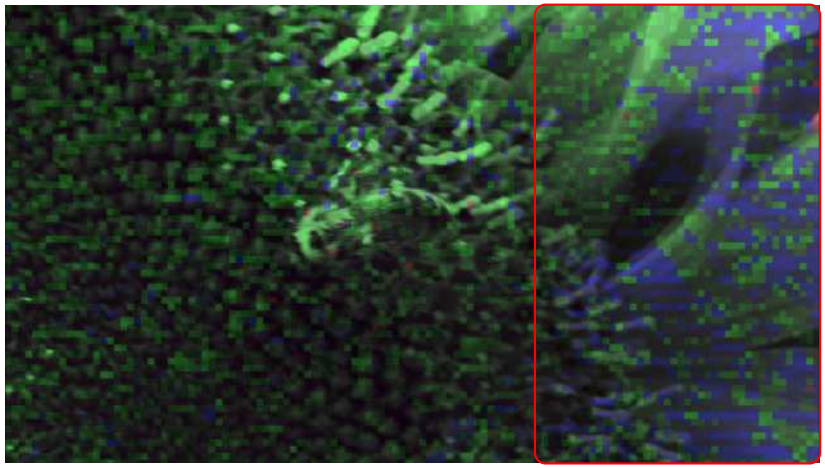
CrowdRun	QP22	QP27	QP32	QP37	Ave.
Total Time	-5.51%	4.80%	10.39%	12.31%	5.50%
Luma-PSNR	-31.17%	-32.74%	-35.15%	-38.56%	-34.40%
Luma-SSIM	-16.96%	-25.68%	-35.48%	-46.88%	-31.25%
Total Bits	-82.97%	-83.33%	-82.34%	-84.37%	-83.25%

Sunflower	QP22	QP27	QP32	QP37	Ave.
Total Time	13.27%	19.29%	20.00%	23.90%	19.12%
Luma-PSNR	-15.10%	-17.51%	-19.29%	-23.63%	-18.88%
Luma-SSIM	-2.80%	-4.80%	-7.88%	-13.88%	-7.34%
Total Bits	-68.86%	-65.60%	-50.29%	-24.49%	-52.31%

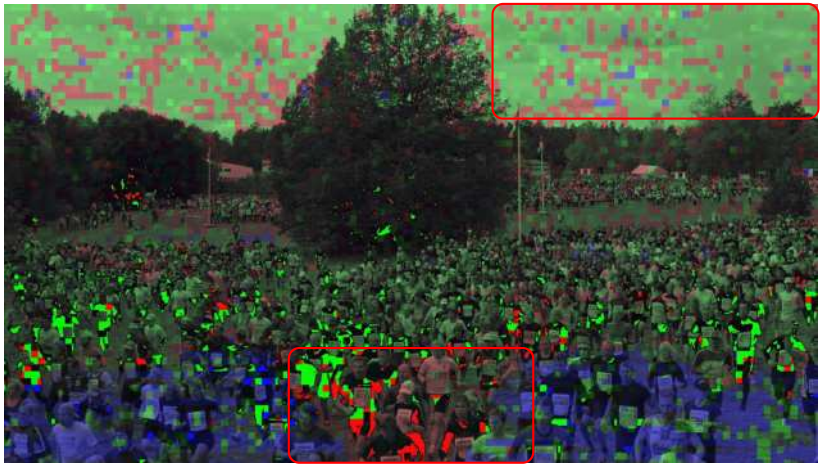
Table 4.3 Summary of LHPSS relative video performance shown as % differences for random access and low delay p prediction structure using CrowdRun and Sunflower 1080p



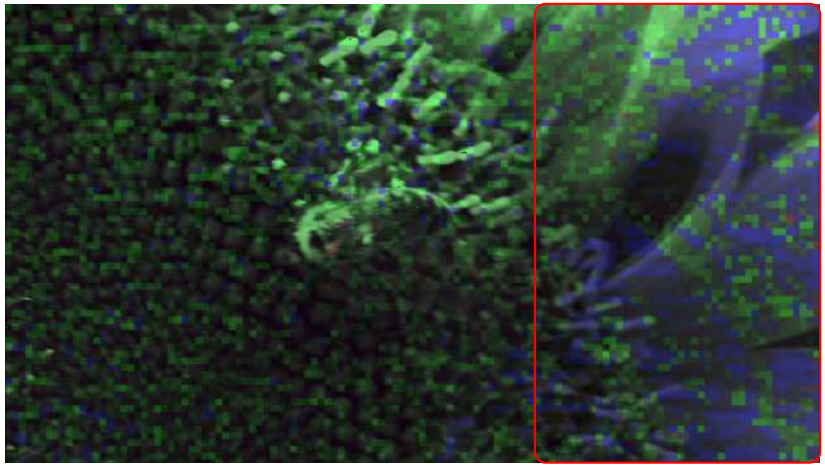
(a) CrowdRun (SATD) I4x4:57, I8x8:6408, I16x16:1695



(b) Sunflower (SATD) I4x4:2910, I8x8:4333, I16x16:917



(c) CrowdRun (Pseudo-SSIM) I4x4:67, I8x8:6161, I16x16:1932



(d) Sunflower (Pseudo-SSIM) I4x4:2127, I8x8:4880, I16x16:1153

Figure 4.13 CrowdRun (left) and Sunflower (right), SATD (top), proposed LHPSS (bottom). Frame 1 (intra) luma where red, green and blue is for number of 4x4, 8x8 and 16x16 blocks respectively.

4.5 Discussion

These results presented here reflect the different experiments. This includes scatter graphs that represent the investigation of the SSIM and STDM relationship as part of the observational study that illustrate the UBR. Considering that existing SSIM based PVC solutions show a simplistic non-linear graph, this provides evidence that SSIM and STDM relationship is more complex. The results from the second experiment extend the knowledge that UBR is based upon covariance by the heat map of the video frame which compares covariance and JND gives support that covariance is perceptual friendly measure. This knowledge that covariance reflect the UBR is used in the construction of the proposed model of the UBR in the form of LHPSS. The LHPSS demonstrates that a HVS friendly changes do occur at the frame level, unfortunately, the overall video sequence image quality and timing performance less undesirable. This is especially disappointing as a broad range of observations were used and also low complexity techniques when producing LHPSS.

4.5.1 Discussion for observational study

This observational study is an experiment to fulfil the lack of evidence for paired scores of SSIM compared to STDMs at the sub-block level. Figure 4.7 is of observations from intra coding, intra frame and shows the response to be limited. Figure 4.8 has a broader range of scores in both perceptual and STDM terms. While Figure 4.9, shows that the video content can affect the response shape. In each case the results will be discussed below.

Single video sequence

Results based upon the single video sequence of Foreman QCIF resolution are shown in Figures 4.7 and 4.8. They are separated by block size and then broken down by distortion assessment. The graphs in Figures 4.7 and 4.8 illustrate the limited response of intra coding compared to the more dynamic response of inter. This can be explained by the way intra and inter coding operate. Similarly, analysing by assessments, the response for the same prediction candidate can differ because of the scales of the distortion metric.

In Figure 4.7 there is a limited response because intra coding has a high likelihood of similarity as they involve adjacent blocks. Also, because of this the statistical properties of the original and reconstructed sub-block are almost identical, meaning the range of 1-SSIM score are limited. In comparison, the results for inter coding in Figure 4.8 show that the response is far broader, as the use of motion vectors can lead to poor block matching in terms of SSIM, STDM or both. Under inter coding it is more likely for prediction candidate having opposing STDM and SSIM scores, one being high and another being low. This indicates that perceptual response is more complex than existing PVC models based on a single non-linear graph and the relationship is a region of shared space.

Overall, this experiment assessed distortion at the prediction stage, where SSE alongside SAD and SATD were calculated together. SSE usually resides in mode decision, however, for this experiment it was added as a means of comparison. The scale used to represent SSE distortion score, is different compared to SAD and SATD because under SSE, the difference is squared. In comparison, SAD and SATD have similarly low distortion metric scores due to the lack of squaring operation, with SATD offering a more proportional and broader response. This means that SATD was chosen as the means to evaluate the second part of the experiment, distortion response across multiple videos.

Across multiple video sequences

This second set of results shown in Figure 4.9 allows comparisons of responses across six different video sequences for both 4x4 and 8x8, with the SATD distortion metric. Each video sequence tested differs in terms of texture and activity, yet throughout, these results occupy a region of shared space. When viewing the observation by their respective video sequences in Figure 4.9 the nature of the video activity exhibits a particular response. As the responses are not lined up along the x-y axis then this shows that the differences due to the relative lighting conditions and this has an effect on the perceptual score. This is shown vividly under Flower 8x8, where SATD score is low, yet the respective 1-SSIM is high, illustrating the perceptual and STDM do not agree at the sub-block level. Figure 4.9 shows in terms of 4x4 compared to 8x8, the smaller block size are limited in range and have limit resemblance in shape to the respective 8x8 graphs.

Looking at the set of graphs by block size shows that they cover a similar distortion metric range. This illustrates that while SSIM fails the $\triangleq$, the shared space could allow SSIM to be compatible. This reinforces the hypothesis and rejects the notion that a non-linear graph is sufficient to represent the 'SSIM-STDM' relationship. From these observations it is possible to infer that a given block size can have its own universal bounded region (UBR). This UBR represents the dimension space between STDM and IQA. The significance of this is that a UBR can be the template to model a perceptual IQA in a STDM space. Overall, being able to model this UBR, can enable perceptually favourable scores for sub-blocks to benefit with increase distortion. This means that the perceptual integrity of the given sub-block is retained for a lower R-D point.

4.5.2 Discussion for identifying components of the UBR

In the graph Figure 4.10 the maximum range for covariance was earlier determined to be $\pm$16k, however, the actual range in these samples were between -334 and 4731. This later lead to the choice of limiting the covariance value to 8000, and this is shown in Figure 4.11.

These findings illustrate that it is possible to assign a given 1-SSIM score to a corresponding SATD value by the covariance of the original and reconstructed sub-block. This means that any model of the UBR can be centred around the covariance of the prediction candidate. Furthermore, covariance in comparison to JND is able to show that homogeneous texture of the foreman's helmet and background panelling have low covariance. This corresponds with the JND version of the same frame, indicating that covariance is liable be higher on textured surfaces, where sub-block matching is difficult. Examples of this include edges of the panel and the foreman's facial features. In all, this means that modelling by way of covariance can allow a perceptual means of scaling SSIM towards STDMs.

4.5.3 Discussion for pseudoSSIM

The results indicate that while medium and larger sub-block sizes are encouraged, the proposed solution has a substantial effect on timing. From the graphical results of Figure 4.13 the reduction of 4x4 usage is clear CrowdRun at the bottom-centre of the frame amongst the crowd. A similar, yet, less pronounce effect occurs

at the top-right of the frame to retain the definition of the clouds. Examining the Sunflower frame indicates that a greater number 16x16 are used, this is highlighted on the right hand side of the frame. This increase of medium and large size blocks refer to SSIM considering the block as a whole than pixel differences individually. As SSIM relies upon statistical properties, this means that candidates with similar properties maybe preferred. At the mode decision where SSE is used, the RDO process collates these new proposed candidates to seek the R-D point closer to the origin. This means there is potential to use larger block sizes or candidates which offer increases in levels of distortion without affecting the perceptual integrity.

In Table 4.3 it shows that for the same video, different results were gained, depending upon the coding structure. For random access, the picture quality changes little, while under low delay p there are substantial losses in both bit usage and picture quality suggests that LHPSS is applied extensively. This means that there is an underlying issue that is causing this discrepancy to occur. This is confirmed when SSIM uses a single windowing operation, which what happens at the sub-block level, making it is liable to select candidates which are statistically similar, when in fact are perceptually dissimilar (Fei et al., 2012). Therefore, while the proposed scaling of SSIM is of low complexity, it is the SSIM calculation which is highly complex. This is especially likely under low delay P where a single reference is used, which can allow more variation to be tolerated. Overall, SSIM is highly complex and its use of a single window for assessment makes it suitable.

4.5.4 Comparing with other existing research

When comparing the results with other existing research, it shows that the experiments presented here go further by gathering observations at the sub-block level and with the statistical properties which constitute the SSIM components (G.-L. Wu et al., 2013; Yeo, H. L. Tan and Y. H. Tan, 2013). Also, the volume of observations across the variety of video content provides a strong basis to extending the understanding SSIM and STDMs relationship as a share-space, UBR, than the non-linear description in existing research (Brooks, X. Zhao and T. Pappas, 2008; Horé and Ziou, 2013). Finally, the pseudo-SSIM solution demonstrates that perceptual at the sub-block prediction stage is beneficial, as it re-orders the candidate selection to retain perceptual clues for highly active content. In existing

research, assumptions are that video content is static or has limited activity, which allow tracking of objects as region of interest or apply perceptual models every key frame (Y.-H. Huang, Ou, Su et al., 2010; T. Huang, S. Dong and Tian, 2014). However, those approaches are liable to struggle when occlusion occurs due to high/dense activity. This means the research presented here of a memory-less sub-block level approach can provide an in-loop solution to retain perceptual integrity without the need to track/monitor objects or frame/block activity. However, the trend towards portable computing means any such in-loop solution should also be low complexity (Chandler, 2013; Su et al., 2012).

4.6 Summary of chapter

Existing SSIM based PVC solutions have been limited to frame or block level due to their λ_p based approach, where an R-Dp curve is modelled and applied retrospectively. This approach restricts SSIM to outside the native sub-block level, however, this is because SSIM does not satisfy the $\triangleq$. This chapter tackled the barrier for SSIM at the sub-block level by introducing a novel means to make SSIM scores STDM compatible. The process revealed that the SSIM and STDM shared a distortion metric space, which was labelled as the UBR. Furthermore, when investigated, the UBR was shown to steerable via covariance, revealing that the UBR reflected a non-linear covariance distribution, with sensitivity to bright areas. This meant that SSIM scores could be scaled up to STDM equivalent scores whist being perceptually aware, reusing existing components. This allowed for a low complexity solution of LHPSS to presented, however, the resulting proposed solution had issues related to timing and image quality. While complexity can be reduced for scaling SSIM to a STDM compatible score, SSIM itself is still highly complex, resulting in a new type of IQA.

Chapter 5

Proposed low complexity pixel based IQAs

Existing PVC solutions are based upon IQAs designed for image coding rather than video coding. This means that popular IQAs like SSIM, which are extensively used in PVC solutions can provide perceptual gain over the reference encoder, at the expense of additional complexity. As explained in the previous chapter, today's application of video coding are moving towards portable and/or low powered devices which have limited processing power. The consequence of this, is that existing PVC solutions are unable to use IQAs like SSIM at the native sub-block level without incurring substantial additional overhead (Su et al., 2012). Techniques have been introduced which can optimise SSIM, at the expense of simplifying the assessment based upon assumptions (Rouse and Hemami, 2008). These approaches are impractical for sub-block level PVC where huge variations in scores are shown in the findings of the SSIM-STDM graphs from previous chapter. Consequently, this means a low complexity IQA is required.

5.1 Related issues with SSIM

The sub-block level experiments from the previous chapter highlighted that the use of SSIM within a PVC solution magnifies the existing issues of complexity and compatibility relative to STDM. The sub-block level is complexity sensitive and SSIM in comparison to other STDMs is complex because of the statistical related calculations. In addition, the arrangement of these statistical calculations result in an incompatibility of SSIM scores with existing STDMs. These two hurdles of complexity and compatibility make the use of SSIM within the sub-block level unattractive, leading to the need for pixel-based IQAs, specific for each STDM norm space.

5.1.1 SSIM and norm space compatibility

SSIM uses windowing to provide an average score, based upon an index. This is not compatible with STDMs which has an accumulated score in either ℓ_1 or ℓ_2 norm space. These different approaches can be made clear when observations with specific scores are analysed as shown in Figure 5.1. In Figure 5.1, specific observations of SSE (ℓ_2), along with their respective SAD (ℓ_1) and SSIM scores are shown. Here observations for where SSE has a score of 200 (green), 2,000 (yellow) and 20,000 (red) only. Under SAD these SSE scores can be described as narrow bands for low (green, SAD 100) and medium (yellow, SAD 250), while for large distortions (red, SAD 500 to 1250) they cover a broad range of scores. Looking by the y-axis for the 1-SSIM scores, it is possible to assign a low 1-SSIM score for all three STDM distortion groups. The graph confirms that SSIM is vulnerable to promote distortion which may have statically similar scores. This supports the discussion in the previous chapter, where SSIM at sub-level performed badly at low delay P. Furthermore, the use of statically similar scores is recognised issue with SSIM based PVC (Fei et al., 2012). The reasons by this behaviour is because the differences maybe high and the windowing process is masking their significance. This means that SSIM may tolerate differences which is both perceptually undetectable and perceptually annoying. For this reason, a new means of IQAs are required, that are able to distinguish effective between perceptually undetectable and perceptually annoying, whilst being STDM compatible.

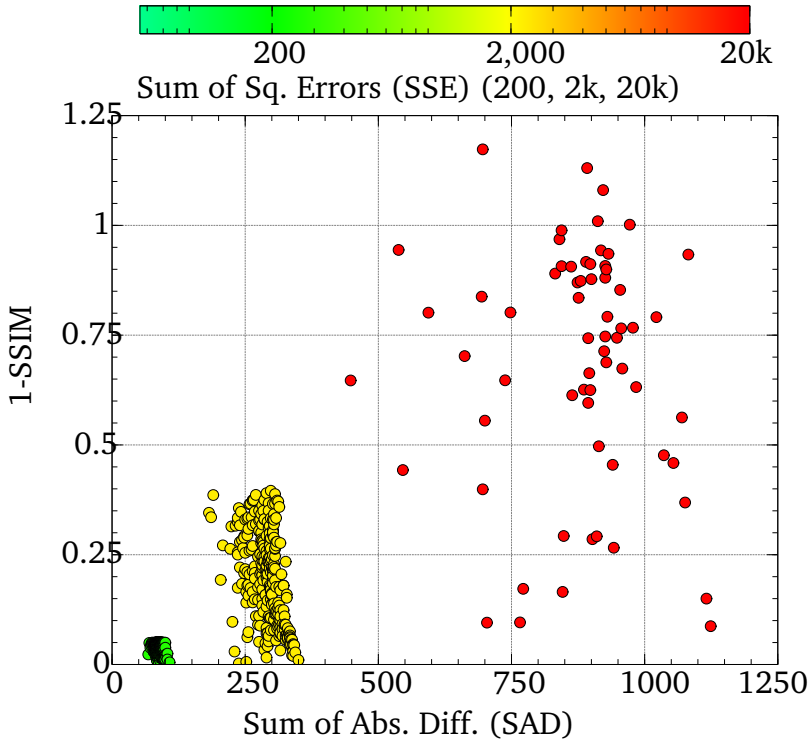


Figure 5.1 Fixed SSE (of 200, 2,000 and 20,000) vs. 1-SSIM

5.1.2 SSIM complexity relative to STDMs

The STDMs used in video coding are SSE, SAD and SATD, which can be categorised into two norm spaces of ℓ_1 and ℓ_2. The norm space provides the distance measures, for SSE where a squared operation is applied, where the norm space category is ℓ_2, while for SAD and SATD it is ℓ_1. These respective STDMS norm spaces can be linked to the level of complexity to calculate the operation. This is shown along with SSIM in Table 5.1 where the number of multiply/divides, addition/subtractions, shifts and absolute functions are required per assessment of an 8x8 sub-block. As shown in Table 5.1, the number of multiples called by SSIM, SSE and SAD or SATD is 208, 64 and none respectively. SSIM is shown as reference, illustrating how its use of statical calculation is highly complex, which reaffirms the need for a low complexity in-loop IQA by others (Su et al., 2012; F. Zhang and Bull, 2015). Overall, this chapter places the challenge to produce IQAs at the pixel level suitable for in-loop PVC while being complexity competitive.

	Multi/Div	Add/Sub	Shifts	Abs
SSIM	208	329	0	0
SSE	64	64	0	0
SATD	0	577	386	3
SAD	0	64	0	64

Table 5.1 Distortion assessment complexity for 8x8 arrays of original and reconstructed.

5.1.3 Need for new pixel-based IQA

In the previous chapter, the experiments used SSIM as the IQA of choice as it has the least complexity amongst its peers, however, it was shown to be too complex for the sub-block level. The issue was further compounded when SSIM and STDM struggled to be reconciled. An example of this issue can be illustrated in Figure 5.1 where low 1-SSIM scores can come across a range of SSE values. This reinforces the other issue experienced during the previous chapter experiments, where under low delay p, where when frames were encoded sequentially, the picture quality for the proposed encoder was very poor. As SSIM operates on statistical properties, then as in Figure 5.1, there is an increased likelihood of candidates being selected with low 1-SSIM and high SSE because under low delay p only a single reference is used. This means that IQA in the form of SSIM is not suitable for video coding, as the use of windowing, averages out changes that could be perceptually significant. Similar concerns have been raised on the performance of SSIM, suggesting dependence on averaging may cause statistics to mask perceptually significant issues (Fei et al., 2012; Zujovic, T. Pappas and Neuhoff, 2013). Therefore, the need for pixel-based IQA is acute and the challenges to resolve issues of complexity are difficult.

5.2 Proposed pixel-based IQA algorithms

Using SSIM as an IQA involves statistical based assessments, which has proved to be computationally costly for the use in sub-block PVC solution. A new approach has been proposed, a pixel-based IQA, which is natively suited to the respective norm space. With this new approach, the additional complexity associated with scaling SSIM to STDM is avoided, as the proposed IQAs should be designed for

video coding environments. Finally, the combined STDM and IQA scores should be compatible and yet be able to influence the candidate choices used by the encoder. This does mean that if an IQA score is applied to a single candidate it must act as a deterrent relative to similar candidates. However, the IQA score should be a viable choice if other candidates choices are less favourable. This is an important principle based upon findings of the previous chapter, where the discovery of a SSIM-STDM shared space led to the modified R-D equation in which κ was introduced to skew the distortion scores.

5.2.1 Need for norm space compatible and low complexity

In order to develop these pixel-based IQAs, an IQA must be aligned to a specific norm space, meaning there will be multiple IQAs. Furthermore, as this will be aimed at the native sub-block level, this means that low complexity is essential. From the previous experiment, it was possible to scale SSIM using low complexity techniques as there was component reuse. This means, that the design of these IQAs must be based upon or at least re-use existing STDM pixel based calculations. While this does limit the design options, however, IQA re-using STDM calculations will benefit from any STDM related hardware acceleration.

5.2.2 Proposed L_2 norm IQA - Sum of Square Differences (SASD)

During mode decision, SSE is used, which is of ℓ_2 norm space. SSE is where the respective pixel difference is squared and accumulated, this means that as the differences increase, the score can rise rapidly. The consequence of this is that an individual difference can have a large impact on the overall score. From an IQA perspective, this may result in different IQA score, yet must be compatible with the ℓ_2 norm space.

The ℓ_2 norm space refers to those values that are squared. Under SSIM, some statistical calculations have compatibility with the ℓ_2 norm space, including covariance which was used in the previous chapter in scaling SSIM to SATD. Using covariance in a pixel based IQA would be ideal, however, the difficulty comes in calculating covariance with respect to low complexity. As covariance is performed using windowing and is processor intensive. Instead, SSE needs adapting to be

perceptually aware and based upon the calculation of covariance. Covariance involves both original x_i and reconstructed y_i pixels, see Equation (5.1).

$$\sigma_{x,y} = \frac{\sum (x_i \cdot y_i) - \Sigma x \cdot \mu_y}{n} \tag{5.1}$$

where $\sigma_{x,y}$ is covariance, μ is mean and n is the block size.

The sum product of original and reconstructed pixels can be rewritten as Equation (5.2), where SSE is $\Sigma(x-y)^2$.

$$\sum (x_i \cdot y_i) = \frac{\Sigma x_i^2 + \Sigma y_i^2 - SSE}{2} \tag{5.2}$$

This allows covariance to be rewritten as Equation (5.3).

$$\sigma_{x,y} = \frac{\frac{\Sigma x_i^2 + \Sigma y_i^2 - SSE}{2} - \Sigma x \cdot \mu_y}{n} \tag{5.3}$$

Unfortunately, Equation (5.3) still involves one multiply and one divide for calculation of covariance, where the divide must exist and cannot be substituted with a right shift. To reduce the complexity further while maintaining the perceptual properties, a new less intensive operation based on Equation (5.2) is required. By adapting the sum of product equation of Equation (5.2) into the difference of squared variables, a measure of sum of squared differences (SASD) is produced in Equation (5.4),

$$SASD = (\left|\left|x_i^2 - y_i^2\right| - SSE\right|)/128 \tag{5.4}$$

The proposed algorithm of SASD in Equation (5.4) has a restricted representation of covariance, in that the squared differences ($x_i^2 - y_i^2$) are used along with the squared sum of differences (SSE). This allows the covariance behaviour of increased sensitivity to higher luma differences to be used, as experienced with previous experiments where covariance was used to perceptually scale scores. While SASD is aimed at utilising ℓ_2 norm space components, it must be downscaled by seven, divided by 128, to ensure it is within a valid range between 0 to 255. As SASD is run on top of SSE, the SASD score should be used to discourage a given candidate than to eliminate it from the list of candidates. The downscaling of SASD

is a crucial stage to limits its influence on ranking of candidates scores compared to others combinations during RDO. For this reason SASD will act as penalty. It is important to see how the proposed IQA of SASD in relation with STDM of SSE, this will be shown visually as response heat maps in Section 5.3.

5.2.3 Proposed L_1 norm IQA - Additional Pixel Cost (APC)

Distortion assessment also occurs at the prediction and rate-control stages with SAD and Hadamard assessments, both of which are ℓ_1 norm space. This does limit what type of perceptual choices are available for use. In terms of IQA, SSIM consists of the SSIM luma equation, see Equation (5.5), which is based upon mean μ, a ℓ_1 norm space compatible component.

$$SSIM_l(x,y) = \frac{2\mu_x \cdot \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{5.5}$$

where C_1 is a constant based upon the maximum pixel range. $C_1 = K \cdot L^2$, where K is 0.01 and L for 8 bit luma is 255 (Z. Wang et al., 2004). The SSIM luma equation has two issues, the first is it has a limited score range of 0 to 1, the second is that it is more tolerant of higher pixel values. Addressing the first issue is possible by multiplying by the bit depth to ensure scores are compatible. In terms of the second issue, SSIM luma can be transformed to be tolerant of darker regions than lighter regions by substituting x and y with $255-x$ and $255-y$ respectively. Unfortunately, the rate of change is limiting, to resolve this, the transformed equation can be accelerated with perceptual model of JND using Equation (5.6), (Chou and Y.-C. Li, 1995; T.-H. Wu, G.-L. Wu and Chien, 2009).

$$JND(x,y) = \begin{cases} 17 \times (1 - (\frac{bg(x,y)}{128})^{\frac{1}{2}}) + 3 & bg(x,y) < 127 \\ \frac{3}{128} \times (bg(x,y) - 127) + 3 & bg(x,y) \geq 127 \end{cases} \tag{5.6}$$

where $(bg(x,y)$ is background luminance, in this case the pixel among original and reconstructed pair with higher value is used. As such, these two perceptual models of SSIM luma and JND can be combined by rearranging and then scaling. The rearranging of the SSIM luma function as described above, makes it in-line with common perceptual principles, labelled as $1-SSIM_l$. Then in order to consider this

as a perceptual cost, it should be scaled by the JND background luminance masking visibility threshold, producing Equation (5.7)

$$APC(x,y) = (2^b - 1) \times (1 - SSIM_l)^{max(JND_x, JND_y)} \tag{5.7}$$

where b is bit-depth and max of JND_x and JND_y refers to using the maximum value of either the corresponding original or reconstructed pixel. In Equation (5.6), JND is a non-linear response curve based on the original frame, while in Equation (5.7) both original and reconstructed pixel values are required to produce a response. This is because the effects of quantisation may change the threshold sensitivity. Similar to SASD, APC will be demonstrated with a response heat map in the next section.

5.2.4 APC on rate-control

In terms of rate-control a special case is required in terms of calculation the final score than each individual pixel. This means that this IQA is not pixel based as such, more like the Hadamard calculation used in rate control, a fixed square block assessment. Hadamard in rate-control applies an 8x8 pixel array from the original frame and treats them as a set differences. This means that only the original pixels are used and under rate-control the assessment calculated is activity and not distortion. As Hadamard is of ℓ_1 norm space, APC is applicable, yet applying

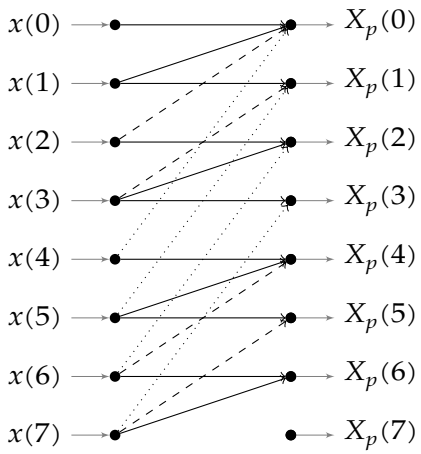


Figure 5.2 Signal flow diagram of proposed weighted perceptual activity pairs based upon Hadamard, where solid, dashed and dotted lines represent weightings of 1/2, 1/16 and 1/64 respectively.

APC in the same way as in prediction would cancel out the eventual APC score. To overcome this issue, APC is applied on the positive pairs, where additions are performed. Also, the pairs are applied with a weighted distance where the distances of 1, 2 or 4, will be downscaled by factors of 1/2, 1/16 and 1/64 respectively as illustrated in Figure 5.2. This process can be referred to as positive pair weighted APC (ppwAPC). The overall effect of ppwAPC is a compressed Hadamard of positive pairs that have been presented, with the last entry of $X_p(7)$ being empty, set to zero. Applying the same process on the y-axis will mean that the last entry in the 8x8 array, the 64^{th} entry will always be zero.

5.3 Compare assessment response heat maps

These proposed algorithms for pixel-based IQA have been designed to reuse existing norm spaces. In the case of SASD, covariance was linked with SSE, which eventually lead to the SASD formula. Similarly, APC is based upon the SSIM luma function as it uses mean, an ℓ_1 norm space compatible component. In either case these proposed algorithms are designed to complement the respective STDMs norm spaces. The 2D response heat map will indicate the behaviour of each respective IQA for any given input. .

5.3.1 Method for comparing proposed pixel-based IQAs

The 2D response heat maps for the proposed pixel-based IQAs involved using a spreadsheet to populate an area of 2^8 by 2^8 (256 by 256) with all possible combinations of the respective IQA and STDMs involved. The choice of 2^8 is because 8-bit video is being considered, which means the range of values are 0 to 255. For the STDMs the process was simple, while for the proposed pixel-based IQAs this involved several spreadsheets to calculate the IQA in stages. The values produced were then used to produce the response graphs using Veusz (Sanders, 2015). Furthermore, to examine compatibility of proposed IQA against STDM, an additional response heat maps were produced to understand what additional downscaling/capping would be required.

5.3.2 SASD response map

To understand how SASD algorithm operates at the pixel layer, three heat maps have been produced, existing SSE, proposed IQA SASD, and the ratio of SASD over

SSE in Figures 5.3 to 5.5 respectively. These graphs are based on 8 bit luma range, 0 to 255 where x and y represent original and reconstructed pixel values. The SSE heat map in Figure 5.3 demonstrates the uniform cost of STDM. In comparison, SASD shown in Figure 5.4 illustrates a non-uniform cost, where the rate of increase in distortion score is higher as pixel values increase. As SASD must be used with SSE, it is worth considering SASD as a ratio of SSE as shown in Figure 5.5. This figure describes a very narrow portion where x is equal to y, this is widening as luma values increase. In Figure 5.5, the spread of values is low relative to the SSE score, this suggests that SSE with SASD will have a limited effect. SSE is used in RDO, mode decision stage, where block matching occurs, which means SASD will be used in influence the sub-block candidate combinations selected. As SASD is similar to covariance, it will be particularly high when attempting to retain the accuracy of high luma values. This presents a major issue, compatibility of SASD with SSE. The response of the ratio of SASD against SSE is virtually zero almost everywhere and four times SSE in a very narrow region, which is far from ideal. To avoid SASD being relatively active compared to SSE, the total SASD score for a sub-block will probably need to be capped relative to the norm space it represents, ℓ_2.

5.3.3 APC response map

The APC algorithm is developed from two perceptual alorithms, SSIM luma and JND. As APC is described as a pixel-based IQA, it is applicable to both prediction and rate-control. Since SAD has a similar linear response to SSE in Figure 5.3, which is why SAD has not been shown and instead, SSIM luma is shown in Figure 5.6. SSIM luma behaves different to common perceptual understanding, in that it provides a higher tolerance for higher luma values. As described, an inverted x and y, equal to 255-x and 255-y respectively is not sufficient. That is where the corresponding maximum JND value is applied resulting in the proposed APC response of Figure 5.7. Therefore, the JND scaling of the inverted SSIM luma function demonstrates an accelerated saturation of APC score as luma values increases, making it sensitive to differences in high values of Luma. When APC is shown as a ratio of SAD, as shown in Figure 5.8, the scales can be larger than that shown in SASD over SSE. However, the extreme scales are limited to top right hand

corner, while most other regions are approximately a factor of 10. Compared to the ratio of SASD over SSE heat map, APC over SAD is even more highly active and probably needs its overall score to be downscaled significantly.

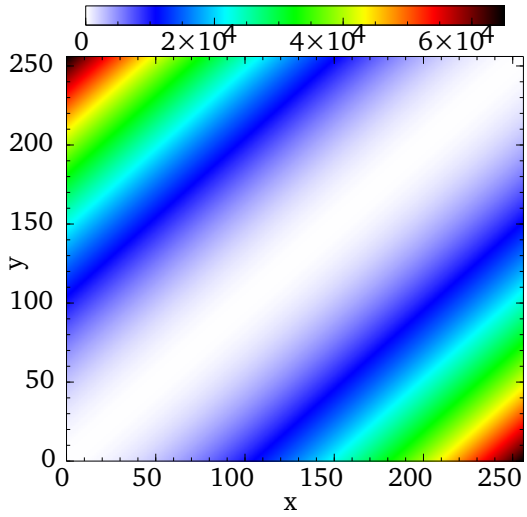


Figure 5.3 SSE Heatmap

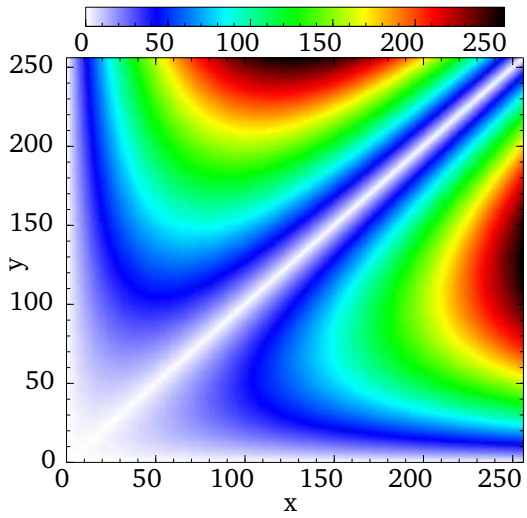


Figure 5.4 Proposed SASD Heatmap

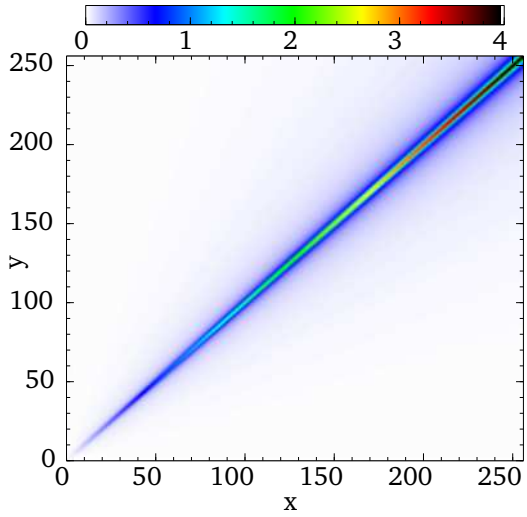


Figure 5.5 Ratio of SASD over SSE

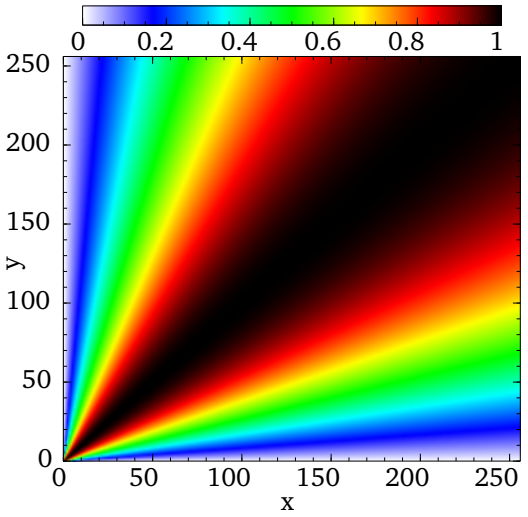


Figure 5.6 SSIM luma equation

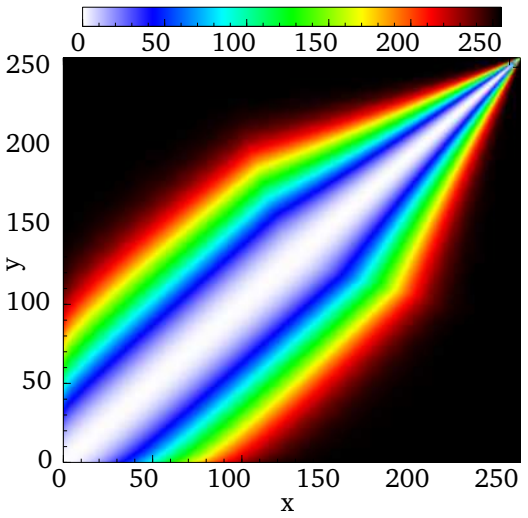


Figure 5.7 Proposed APC

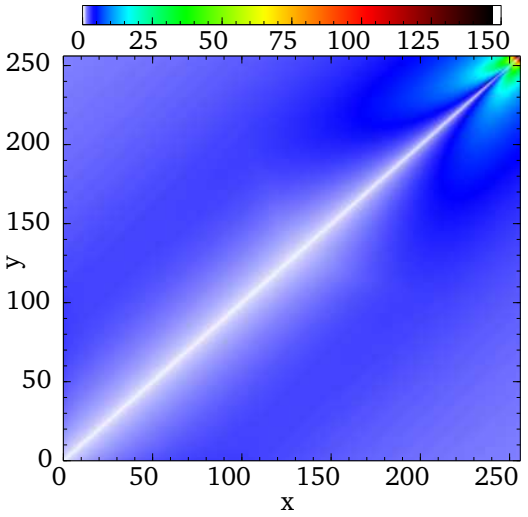


Figure 5.8 Ratio of APC over SAD

5.4 Methodology for pixel-based IQAs and visual simulation

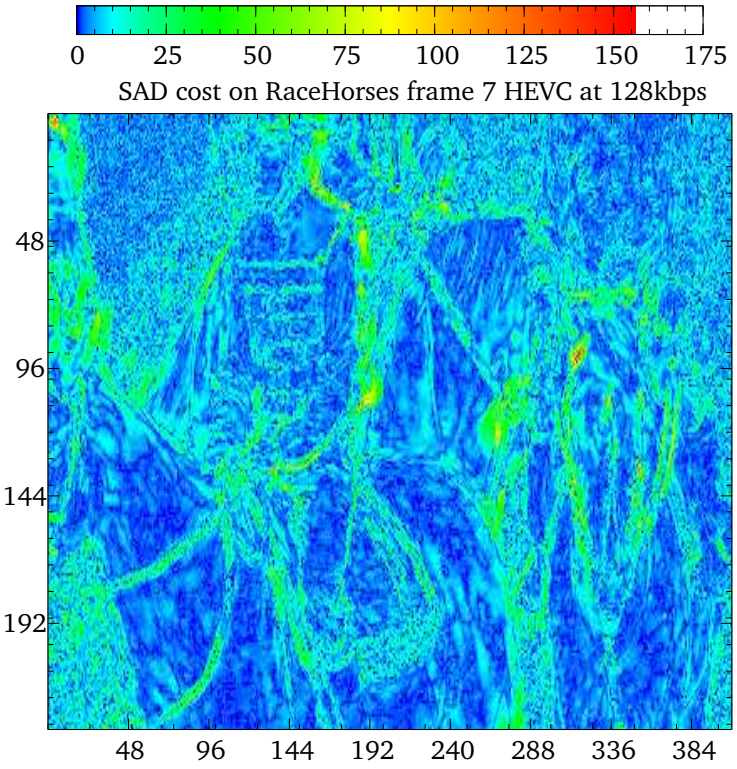
To evaluate the proposed pixel-based IQAs, the assessment was applied to a video frame from an encoded video sequence. In this case a frame 7 from the video sequence called RaceHorses (416x240 pixels), encoded using HEVC at 128kbps under random access profile. Then the decoded and original luma values were used to calculate the STDM and proposed IQA for each pixel, except for the last 8 pixels on the width and height of the frame. Once the numerical values had been logged, the graphing tool called Veusz was used to highlight the effects of using the IQA in isolation, without the low complexity framework.

5.5 Results for IQA and STDM on video frame

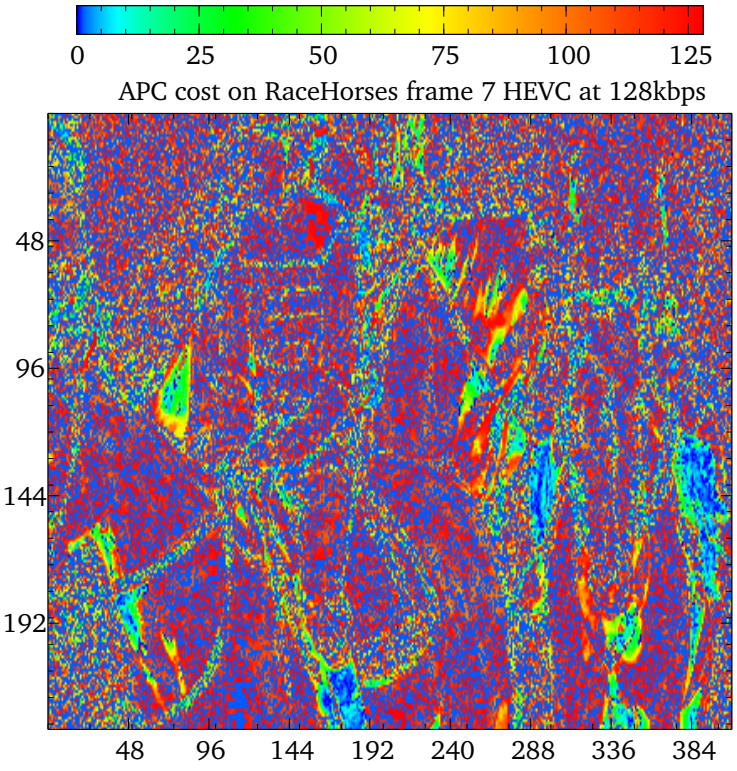
The results presented here are IQA on video frame. The heat maps demonstrate the proposed pixel-based IQA behaviour irrespective of the thresholds applied within the hybrid STDM-IQA workflow framework. This means that these heat maps reflect how pixel-based IQA natively perform prior to any scaling. In Figures 5.9a, 5.9b, 5.10a and 5.10b the distortion from the HEVC encoded video frame is assessed using SAD, APC, SSE and SASD respectively. The assessment is applied to every pixel except for the last 8 pixels of the height and width of the frame. The heat map colour scheme where blue is low no distortion, to red high distortion.

5.6 Discussion of results

When the proposed IQAs were applied to video frames as shown in Figures 5.9 and 5.10, the proposed IQAs demonstrated relatively more activity than the equivalent STDM. APC scores are more dynamically spread compared to SAD, suggesting that APC should be applied selectively, as it assigns cost to virtually all distortion. This reinforces the need to have a pre-checks for APC before it is applied to a sub-block as an additional cost. While for SASD, it also has a higher dynamic activity than SSE, this is shown particularly in the background and on the horses. Gathering these results is time-consuming, and has a high chance of human

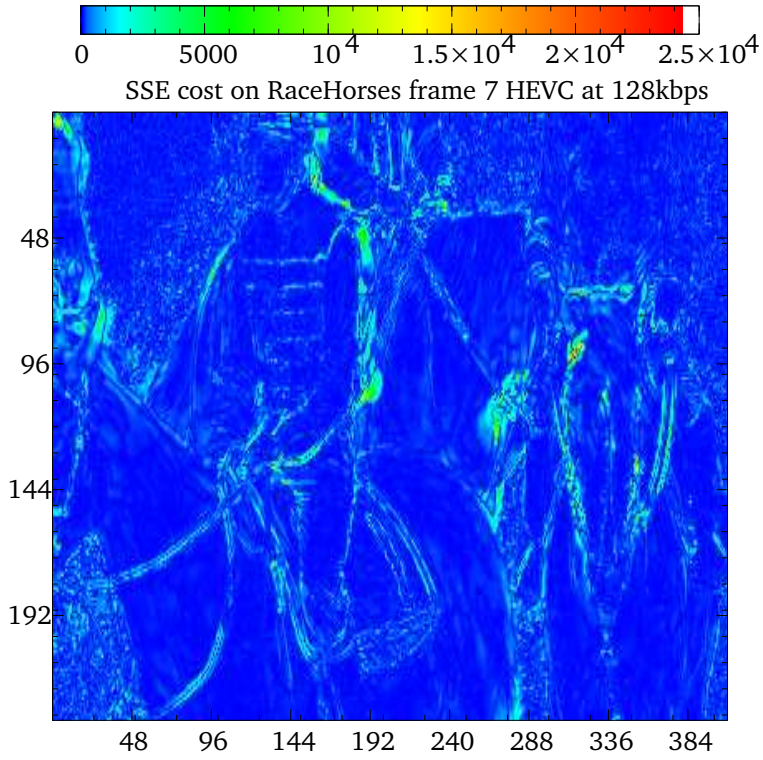


(a) SAD heat map of frame 7 from RaceHorses encoded at 128kbps

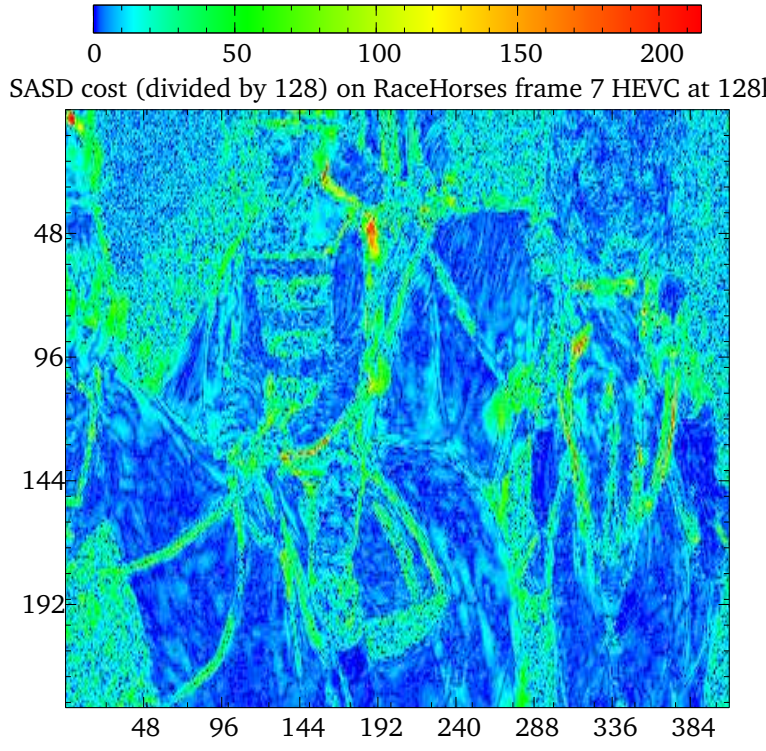


(b) Proposed APC heat map of frame 7 from RaceHorses encoded at 128kbps

Figure 5.9 ℓ_1 norm STDM and IQA, SAD and APC heat map of frame 7 from RaceHorses encoded at 128kbps



(a) SSE heat map of frame 7 from RaceHorses encoded at 128kbps



(b) Proposed SASD heat map of frame 7 from RaceHorses encoded at 128kbps

Figure 5.10 ℓ_2 norm STDM and IQA, SSE and SASD heat map of frame 7 from RaceHorses encoded at 128kbps

error, consequently, there is a need to automate this process via a tool which can also aid the development of IQAs. As such rate-control is not shown as it is complex, this will be tested in the next chapter within a simulation of the proposed encoding process. However, the proposed ppwAPC IQA for rate-control shares similar traits to the proposed APC IQA for prediction and the results presented here of APC should provide an indication of its behaviour. In either case, the these IQAs can not be applied alone, instead they should operated in conjunction with the respective STDMs on a conditional basis.

5.7 Summary of chapter

From the literature review video coding continues to grow in applications for low powered and/or portable devices. The initial set of experiments in the previous chapter were able to reduce the complexity associated with scaling of SSIM to be compatible with existing STDM, yet this highlighted the high complexity of SSIM. This chapter presented video coding in mobile and non-traditional environments, meaning that low complexity PVC is requires should use IQAs which are pixel based than window based used by SSIM. In turn, this led to the design and development of pixel-based IQAs centred around the respective norm spaces of the front-end distortion and activity assessments stages. The results showed that these new IQAs are perceptually aware, however, they seem over highly active will need to be capped and/or downscaled to be used in the respective norm spaces. To stabilise this, it will be beneficial to apply these pixel-based IQAs with existing STMDs, however, this raises questions of knowing when and by how much. As such, these questions forms the basis for the investigation for the next chapter, in finding a means to utilise low complexity pixel based IQAs to be part of a STDM-IQA framework.

Chapter 6

Proposed STDM-IQA framework

As shown in the earlier chapters, using SSIM alongside existing STDMs have issues of compatibility. This means that the approach used by SSIM of windowing is impractical. This led to the development of pixel based IQAs, designed with reduced complexity than using SSIM and have compatibility with the respective STDM. However, pixel-based IQAs can be highly responsive and so should be used in with STDM only where distortion or activity is perceptually significant. To ensure a complexity competitive solution, these IQAs should be called as required, subsequently, it is a case of knowing when to apply IQAs and thus producing a hybrid STDM-IQA framework. In order to develop and test this new framework, it will be modelled on captured data and simulated on a proposed visual VCL tool to understand which observations are affected and where it triggers on the frame respectively. Therefore, this chapter is about designing a complexity competitive PVC solution based upon a proposed a hybrid STDM-IQA framework suitable for where IQAs are called when distortion or activity is perceptually significant.

6.1 Proposed hybrid STDM-IQA framework

During the early part of the previous chapter in Table 5.1, it had been shown that existing IQAs of SSIM are highly complex relative to STDMs, and are recognised as a major issue (Chandler, 2013; F. Zhang and Bull, 2015). This is because STDMs have a linear response within their respective norm spaces, which make them highly efficient. In comparison, an IQA applies a non-linear response, however, according to Weber's law there is a common initial stage which perceptual and non-perceptual response is similar (Peli, 2001). Under JND, this is described as the minimum threshold at which variation would be perceptually detectable for a given luma intensity, as shown in Figure 6.1. This means perceptual sensitivity varies depending upon lighting, therefore, the issue can be re-framed as applying IQA where relevant. In particular, the region below the perceptual threshold can be interpreted with linear response of an STDM and those at or above with a non-linear response of an IQA. Under this premise, using an IQA is an additional assessment to STDM, when conditions are met. Consequently, an STDM may or may not call

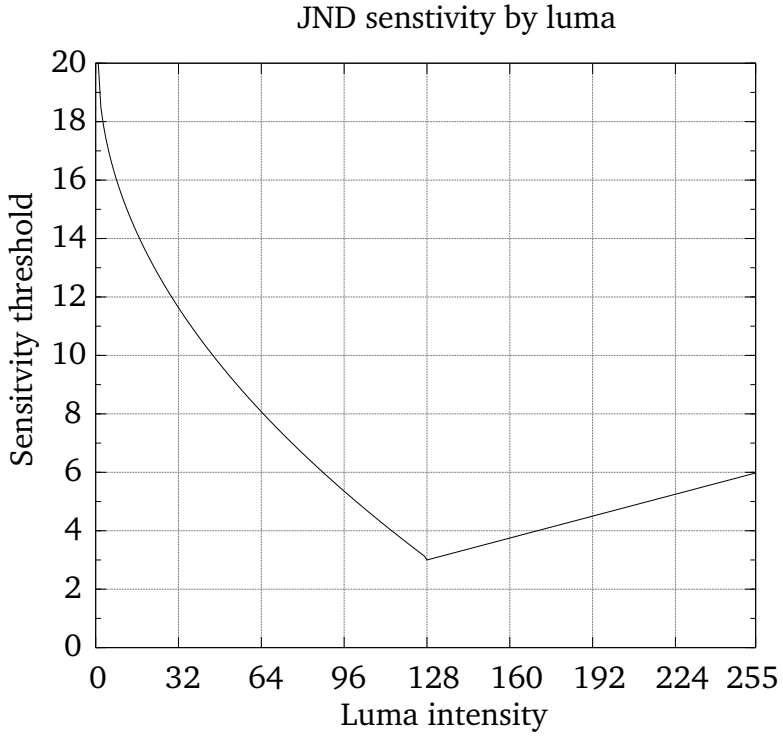


Figure 6.1 JND sensitivity threshold by Luma

an IQA, and this decision making process is a critical stage. This means that it is crucial to evaluate where candidates require the additional IQA assessment.

6.1.1 Low complexity proposed framework

To ensure low complexity, this means any evaluation to decide whether an IQA should called must operate upon a sample of sub-block. That way, any further operations as part of the framework is minimised, especially any evaluation for perceptual significance will be applied to all candidates. In all, the design of this concept can be illustrated in Figure 6.2, where the proposed hybrid STDM-IQA framework is presented.

In order for this proposed hybrid STDM-IQA framework to be made possible two parts are required, first a means to evaluate whether a sub-block is perceptually significant, and second a pixel-based IQA score. The former, a means to evaluate perceptual significance is critical as it is a complexity saving measure. However, this does depend upon the test process being less complex than the proposed IQA. For the later, new IQAs are required, design to be pixel-based and compatible with the respective norm spaces. Overall, this will allow for perceptual scores to be added without the need to scale using complex means.

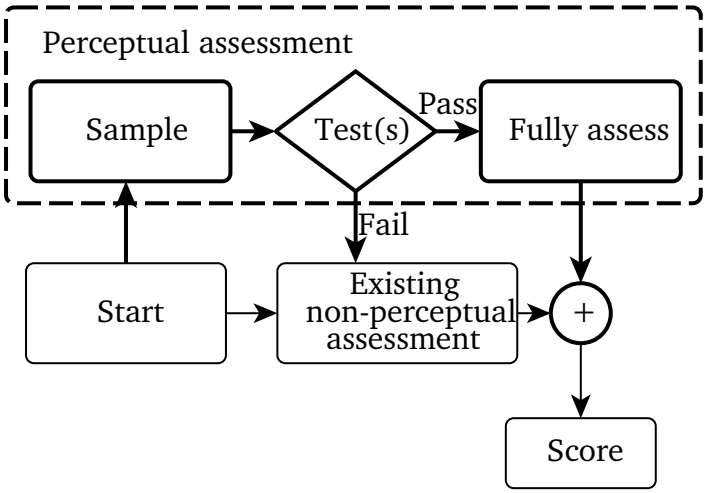


Figure 6.2 Proposed hybrid STDM-IQA framework

6.1.2 Argument for combined assessment over single perceptual assessment

Ideally, it would be convenient to have a single perceptual metric score than an STDM plus an IQA score. As described in Section 3.5.3, a true distortion metric supports the $\triangleq$, while as demonstrated in Section 4.1.3, SSIM, an existing perceptual assessment fails to satisfy the $\triangleq$. This is because the non-linear cost of perceptual assessment is liable to exceed the $\triangleq$. With SSIM, as shown in Table 4.1, there was >6.7k% difference when testing for the $\triangleq$, while for SATD which supports the $\triangleq$, there is a difference of $\approx$ 1%. Initially, the approach for single IQA replacement solution for STDM led to the discovery of the UBR, a region of shared space representing the relationship SSIM and SATD. It was even hypothesised that the UBR could be manipulated, which was represented by κ, while the UBR was mapped to scale SSIM to produce pseudo-SSIM, κ was not implemented, However, by applying IQAs on top of STDMs, this suggests that an IQA score will act as κ and break the $\triangleq$ for the respective STDM.

Proportion of IQA score relative to the STDM

Use of the proposed framework with the IQA path will result with an addition distortion score for the STDM, however, what proportion of IQA relative to the STDM is undefined. Using Figure 5.1 as a guide, choosing fixed SSEs of 200, 2,000 or 20,000 each have variation that is significant against SSIM, increasing as SSE score increases. This may be related to SSIM applying windowing, and STDM being pixel based, however, the same description can all be applied when comparing the two STDMs, as they are based upon different norm spaces. Therefore, a cautious approach would be to set the level of proportion of IQA score relative to STDM to $\approx$ 10%. Overall, this figure is a starting point and will be reviewed by modelling and simulating the framework to understand the potential effects on the frame/video content.

6.2 Perceptual significance tests for proposed hybrid STDM-IQA framework

Underlying the proposed hybrid STDM-IQA framework concept is the use of perceptual significance tests. These tests act as a means to judge whether the distortion or activity is of perceptual significance based upon a sample of pixels. Consequently, these tests are based upon the assumption, that a given sample is reflective of the sub-block as a whole. This is because the need for the tests to be less complex than the assessment offers a potential complexity savings technique. However, to encourage robustness of whether distortion or activity is of perceptual significance, different types of tests are applied. These mean using pixel samples that include comparing both opposing and adjacent pixels to measure the rate of change within a block or across a pixel.

6.2.1 Proposed perceptual asymmetrical side (PAS) test

Sampling the sub-block sides allows the broadest possible spread to be tested, because where the perceptual difference is homogeneous, the sides will be identical. This is illustrated in Figure 6.3 where the sub-block sides, minus the corners are highlighted. If these sides represent a homogeneous set of perceptual differences, then a uniform cost of an STDM will be more suited. Consequently, it is where these perceptual differences are non-homogeneous that they may represent a sub-block candidate that is of perceptual interest. Underlying this is the assumption, that the sub-block sides is reflective of the inner square. Another assumption is

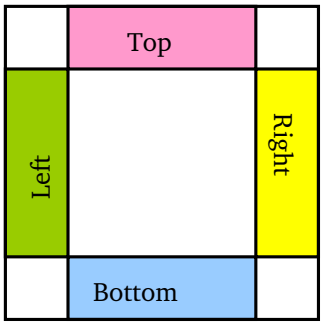


Figure 6.3 Perceptual significance test

that a perceptually homogeneous sub-block is where the sub-block sides are similar, meaning that opposing sides can cancel out. Therefore, where one side is dominant, the sub-block candidate is of perceptual interest. From these assumptions an equation can be produced to test perceptual significance of an 8x8 block. Here, a proposed test is where each side of the perceptually normalised version of the sub-block is taken (minus the corners) and subtracted from each other to find perceptually asymmetrical side labelled as PAS in Equation (6.1),

$$PAS = (||T + B| - |L + R||) > PAS\ Threshold \tag{6.1}$$

where T, B, L, R and $Thresh$ are top, bottom, left, right and threshold respectively. These sides will be assessed using the respective IQA, for ℓ_1 or ℓ_2 norm space. This means that the threshold is different for SASD and APC based PAS.

6.2.2 Proposed APC cross corner subtraction (ACCS)

As described above, PAS on prediction is potentially rather costly due to the high volume of candidates. Consequently, this would impact the complexity for encoding video, therefore, a simplified version is required for prediction to support the variety of sub-block sizes. Under HEVC the LCU can be up to 64x64, also this can include asymmetric sub-block sizes (where width or height of sub-blocks are 12, 24 or 48). This requires a solution that can evaluate sub-blocks for their potential perceptual significance faster with assessing perceptual distortion. As mentioned in the framework, this means evaluate a sample of the sub-block first to judge whether an IQA be applied. For mode decision PAS is proposed where opposite sides are evaluated, while for prediction it is proposed that opposing corners be evaluated to reduce the level of processing required. Using the corners to evaluate whether an IQA should be used means that stabilisation of averaging sides used in PAS is not available. Instead, the difference of differences is used, similar to finding patterns in number sequences or to measure acceleration, which in this case the cross pair differences are shown in Equation (6.2).

$$ACCS = |(A_{TL} - B_{BR}) - (C_{TR} - D_{BL})| > ACCS\ Threshold \tag{6.2}$$

where A_{TL}, B_{BR}, C_{TR} and D_{BL}, denote the sub-block corners, top left, bottom right, top right and bottom left respectively. While ACCS means APC cross calculation subtraction, where the respective diagonal corners are subtracted from each other based on their APC values. The ACCS threshold governs at whether IQA should be performed, this setting will be discussed later. The equation was designed to be low complexity, using only a single absolute function.

The use of ACCS on the block corners is limited to four test points and was designed under an 8x8 sub-block. Therefore, every whole multiple of 8x8 should undergo a further process of ACCS to ensure that add a level of robustness. This can be illustrated with Figure 6.4 which shows that the two stage process as outer, the sub-block corners, and as inner, for every 8x8 within.

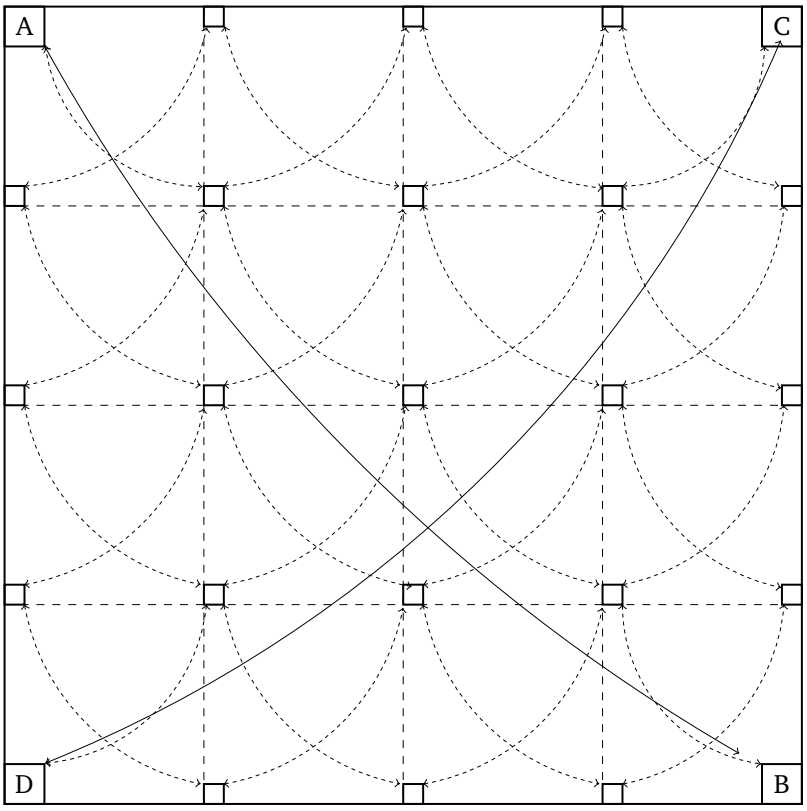


Figure 6.4 APC cross calculation subtraction (ACCS) outer and inner (applicable to >8x8 sub-blocks)

6.3 Design for perceptual significance threshold

These perceptual related tests and the encoder specific operations are shown to depend upon thresholds to indicate whether additional perceptual processing is required. PAS is applied to mode decision and rate-control, using the proposed SASD and ppwAPC algorithms respectively. This allows eliminating candidates that are perceptually homogeneous or symmetrical, and focus on capturing candidates which exhibit perceptually significant uneven distribution. While a prediction specific set of thresholds are presented for use with ACCS.

6.3.1 Use of non-linear scaling of thresholds

The use of thresholds is to apply IQAs on those candidates with perceptually significant distortion or activity. While this is important to apply to each candidate within each sub-block, this process would continue in each sub-block sizes. This means the IQA cost could be applied to multiple smaller sub-blocks against larger sub-blocks at the same time, which when RDO, mode decision occurs could revert existing way of choosing smaller sub-blocks. Therefore, to discourage this practise within the encoder, the threshold should be scaled non-linearly, that way the use of smaller blocks is applied where distortion from larger sub-blocks is great. This means that initial threshold must be based upon sub-block size which is suitable to scale from. Within the HEVC encoder 8x8 offers substantial number of observations and is also recognised as a stable sub-block size (Brooks, X. Zhao and T. Pappas, 2008). Overall, this would allow perceptually significant regions to have smaller sub-blocks allocated, while larger sub-blocks would be placed in perceptually homogeneous regions, thus encouraging bit re-distribution.

6.3.2 Perceptual side and corner thresholds

PAS is suitable for where there is sufficient time to gather perceptual differences on the sub-block sides, this is possible for rate-contol and mode decision, and unsuitable for prediction due to the volume of candidates. This means that PAS is aimed at two different norm spaces, therefore, two separate thresholds need to be determined. The encoder was modified to log observations during prediction Hadamard 8x8 to calculate PAS under SASD and ppwAPC. However, upon simulation these initial thresholds were revised down to binary friendly

values of 48 and 8 for PAS in mode decision and rate-control respectively. This was decided based upon the simulation via the proposed visual VCL tool, which highlighted where on the frame the IQA would potentially be applied.

6.3.3 Mode decision: SASD linear and non-linear thresholds

Compared to prediction, in mode decision there are a limited number of candidate combinations evaluated by the RDO process. As the IQA of SASD called conditionally when pre-tests of PAS are met, the resulting SASD may be insignificant and even potentially detrimental. Thus, SASD should only be added to SSE if a post check SASD threshold has been met.

To establish a threshold the encoder was first modified to capture observations, which were then analysed. The Hadamard 8x8 assessed was modified to extract both original and reconstructed data. This resulted in a total of over 1 million captured observations based upon the first 3 frames of RaceHorses video sequenced encoded in HEVC at 1024kbps. Using R, 330,000 unique observations were identified and a sample of 30,000 was then used. The descriptive statistics stated that the median was SASD of 496, and this corresponded to observations which had higher proportion of 1-SSIM values. As the purpose of SASD is to discourage perceptually annoying distortion, the median value was deemed satisfactory setting for this purpose. However, the threshold was set to 512 for 8x8 sub-blocks as it is a binary friendly number. From this, the post-check threshold for SASD was scaled linearly for other sub-block sizes, as described in Equation (6.3).

$$SASD\ Threshold_{Linear} = 2^{(2n+3)} \tag{6.3}$$

	SASD Threshold				
Block size	Linear	log_2	Non lin	log_2	Factor
4x4	128	7	128	7.00	1
8x8	512	9	768	9.58	1.5
16x16	2048	11	3584	11.81	1.75
32x32	8192	13	15360	13.91	1.875
64x64	32768	15	65536	16.00	2

Table 6.1 Mode decision: block size non-linear threshold values for SASD

where n is $log_2(blockwidth)$, with the respective threshold per block size.

Initially, the SASD thresholds were designed with a linear scaling, this places uniform weighting irrespective of block size. However, larger sub-blocks should be encouraged in perceptually homogeneous regions meaning a non-linear threshold is required, which would be more discerning of when to use smaller blocks sizes. This can be established by Equation (6.3) as basis to produce non-linear threshold as shown in Table 6.1. Under the non-linear response, the rate of change rises higher than shown by the factor column, relative to the linear thresholds, eventually doubling the threshold when block size is 64x64. The rate of change was based upon the non-linear factor multiple was growing similar to a log function, accelerating and then saturating at 2.

ACCS Threshold (non-linear scaling by block size)										
Block Height	Width (Thresh)	4	6	8	12	16	24	32	48	64
4		176	200	216	240	256	272	288	312	328
6		200	224	240	256	272	296	312	328	344
8		216	240	256	272	288	312	328	344	360
12		240	256	272	296	312	328	344	368	384
16		256	272	288	312	328	344	360	384	400
24		272	296	312	328	344	368	384	408	416
32		288	312	328	344	360	384	400	416	432
48		312	328	344	368	384	408	416	440	456
64		328	344	360	384	400	416	432	456	472

ACCS Threshold (as % relative to 8x8 reference)										
Block Height	Width (%)	4	6	8	12	16	24	32	48	64
4		0.69	0.78	0.84	0.94	1.00	1.06	1.13	1.22	1.28
6		0.78	0.88	0.94	1.00	1.06	1.16	1.22	1.28	1.34
8		0.84	0.94	1.00	1.06	1.13	1.22	1.28	1.34	1.41
12		0.94	1.00	1.06	1.16	1.22	1.28	1.34	1.44	1.50
16		1.00	1.06	1.13	1.22	1.28	1.34	1.41	1.50	1.56
24		1.06	1.16	1.22	1.28	1.34	1.44	1.50	1.59	1.63
32		1.13	1.22	1.28	1.34	1.41	1.50	1.56	1.63	1.69
48		1.22	1.28	1.34	1.44	1.50	1.59	1.63	1.72	1.78
64		1.28	1.34	1.41	1.50	1.56	1.63	1.69	1.78	1.84

Table 6.2 Non-linear threshold and percentage equivalent (where 1.00 is 100%) (with reference to 8x8) for APC cross corner subtraction (ACCS).

6.3.4 Design for ACCS threshold in prediction

For prediction the same raw luma observations from an 8x8 sub-blocks were used to determine the ACCS threshold. This choice of 8x8 over other sub-block sizes was because a smaller sub-block size would potentially be influenced by a few pixel differences, as shown investigating for the optimal window size for SSIM (Brooks, X. Zhao and T. Pappas, 2008). While a larger sub-block value could equally mask trends and the number of observations would be dramatically less. Equally, because of this stability and number of observation, this 8x8 threshold would be scaled and applied to other sub-block sizes.

Initially, the threshold value was set to 128, this was based upon descriptive statistics using observations gathered from the first three frames of CrowdRun, a HD video sequence. However, because a high number of false triggers were observed during visual simulation using the proposed visual VCL tool, where BasketballDrive was used, the threshold was doubled to 256. In addition, to extend this process for other sub-block sizes a non-linear scaling was formulated, which rounded numbers to the nearest 8th. This is expressed as a formula in Equation (6.4).

$$ACCS\ Threshold = (Int(log_{128}(2 \cdot blksize) \cdot 32)) \cdot 8 \tag{6.4}$$

where blksize is the size of the sub-block, width times height. Using Equation (6.4), Table 6.2 was produced showing the thresholds for the various combinations of sub-block sizes in HEVC. Overall, the formula and table illustrate a non-linear scaling with smaller sub-blocks having a relative lower threshold compared to larger blocks. This way, smaller sub-blocks are more likely to have IQA scores associated and thus during RDO smaller blocks are used to preserve where candidates are liable to produce perceptually significant distortion.

6.4 Proposed perceptual edge detect

The use of perceptual side and corner thresholds can potentially identify those with non-homogeneous spread. However, it is simplistic in that it depends upon a fixed threshold, meaning another type of perceptual test is required to increase

robustness. While perceptual side and corner thresholds are based on averaged or single pixels, they consider change across the sub-block as a whole than by adjacent pixels. This means that sudden changes, such as boundaries, textures and in lighting which represent perceptually significant features are not detected. It is possible to observe this local change by applying edge detection. However, current forms of edge detection require large number of entries and is liable to false triggers. As most edge detections are large and cumbersome, a new proposed edge detection is presented in the form of a 2x2 sized edge detection as shown in Equation (6.5) and in Figure 6.5,

$$Edge \Leftrightarrow (2 \cdot Centre) > (Top + Left) \tag{6.5}$$

where *Edge* is true if the conditions are met, with *Centre*, *Top* and *Left* are pixel values.

6.4.1 Edge detection on rate-control

The proposed edge detection can be adapted for the 8x8 sub-block in rate-control by assessing the block corners using the respective sub-block side average values. As explained earlier, ppwAPC is the perceptual IQA used on rate-control, since edge detection uses the perceptual IQA values, each respective pixel must undergo ppwAPC, however, the last pixel of an 8x8 ppwAPC has no transform. This means that edge detection on the bottom right hand corner of the 8x8 array will always return false, which leaves three corners that can be tested for perceptual significance within the sub-block. Thus, only when there is a majority of edges detected, minimum of two, will the perceptual score based upon ppwAPC be calculated.

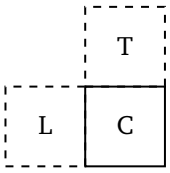


Figure 6.5 2x2 edge detect, where T is top, L is left and C is centre

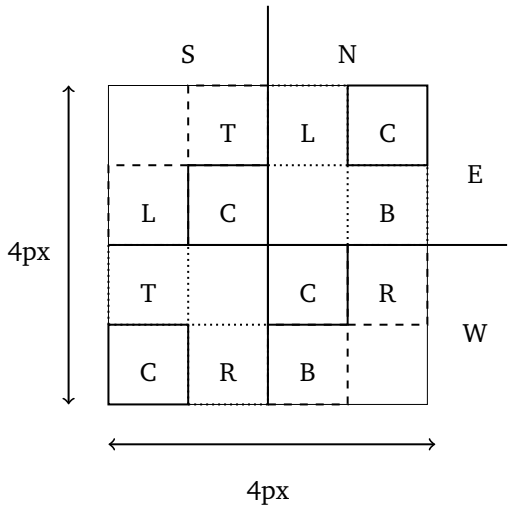


Figure 6.6 2x2 edge detect in different orientations (clockwise, NE, NW, SW and SE), where T is top, B is bottom, L is left, R is right and C is centre

6.4.2 Edge detection for mode decision

In mode decision, RDO is the last assessment stage and this means that perceptual integrity of the sub-block as a whole is important. For this reason the edge detection is applied uniformly in non-overlapping manner across the sub-block. This is shown in Figure 6.6, where for every 4x4 pixels the edge detection is applied in different orientations, north-east, north-west, south-west and south-east. This pattern can be described as windmill-compass, however, this means that edge detection coverage in mode decision is limited to a maximum of 1/4 block size pixels.

The threshold for edge detection in rate-control, was set to be the majority of edge test point. For mode decision this majority threshold continues, where the edge detection threshold for all block sizes, the number of test points required to have an edge detect is greater than half. As mode decision has many tests points, this does mean that along with the majority threshold, this approach should minimise the risk of candidates being processed with SASD IQA based upon phantom edges which act as false positives. This fixed threshold of half can be represented as an equation in Equation (6.6)

$$Edge\ count = 2^{(2n-3)} \tag{6.6}$$

Similarly, a non-linear threshold for edge detection can be applied depending upon block size to encourage IQA on smaller blocks as opposed to larger blocks. This is shown in Table 6.3 where the edge threshold increases from 1/2 to 3/4 from 4x4 to 64x64 sub-block sizes.

6.4.3 Edge detection for prediction

Prediction is the initial assessment stage and where less scrutiny can be applied due to the volume of candidates. For this reason, edge detection is placed on sub-block corners only, yet the orientation of the edge detection will vary, similar to that shown in Figure 6.6. Overall, using Figure 6.4 as a guide, the four corners of 'A', 'C', 'D' and 'B', would have edge detection orientated as south-east, north-east, south-west and north-west respectively. This means that unlike mode decision where multiple test points are used, which increase with block size, a fixed number of four are used in prediction. For prediction edge detection threshold is fixed at 1/2, or minimum of 2 test points must state an edge has been detected to signify the sub-block is of perceptual significance. This threshold of 2 was shown to be sufficient when 32k observations for each bit rates of 1, 4 and 16 Mbps were captured from the decoder and analysed in R. By modifying the decoder only the best candidate is available which does reduce the choice of candidates, while making the data manageable. The data gathered was simulated in R, to examine the effects of thresholding. With a setting of 2, this reduced ACCS based observations by $\approx 1/2$ (down to 52%), this is significant and allows subsequent stage of APC to be applied selectively. However, this opens the question of which areas of the frame are affected and this will be explored later by the proposed visual VCL tool in Section 6.8.

	Edge Threshold (No. of edges must be greater than)						
Block size	Linear	log_2	% of Max	Non lin	log_2	% of Max	Factor
4x4	2	1	0.5	2	1.00	0.5	1
8x8	8	3	0.5	10	3.32	0.625	1.25
16x16	32	5	0.5	44	5.46	0.6875	1.375
32x32	128	7	0.5	184	7.52	0.71875	1.4375
64x64	512	9	0.5	768	9.58	0.75	1.5

Table 6.3 Mode decision: block size non-linear threshold values for edge detection

6.5 Complexity of workflows

At this stage, the tests and thresholds have been defined for STDM-IQA framework design for each norm space respectively. The complexity of each proposed hybrid STDM-IQA workflow involved producing calculations at each stage, of pre-check(s), IQA and post-check (where applicable). This involves going through the codebase for the implemented STDM-IQA workflow and ensuring it followed the designated IQA path. For mode decision, the additional processing is related to the block size, however, this allows the complexity to be scaled by block size. While for prediction despite having multiple block sizes, the additional cost was relatively low and fixed, except for where block sizes > 8x8. Finally, in rate-control, a single fixed block size of 8x8 is applied throughout the frame, meaning that absolute numbers can be shown.

In these set of tables (Tables 6.4 to 6.6) and graph (Figure 6.7) the proposed IQAs are part of the hybrid STDM-IQA framework, where the IQA workflow path is undertaken. This means that each can be described as a variation of pre-check(s), IQA, post check. Since these IQA workflows are designed to be of low complexity these tables and graph provide a breakdown of the proposed stages or mathematical operations required to perform each IQA workflow path.

6.5.1 Overview of proposed IQA workflows

At this point it is important to understand what the proposed hybrid STDM-IQA framework will consist of for each front-end encoder structure. The proposed hybrid STDM-IQA framework revolves around having an overall low complexity, thus the design is presented as a conditional IQA. For this conditional IQA to occur, perceptual significance tests are undertaken, as pre-check(s) and/or as a post check. Combining the proposed perceptual significance tests and assessments with their respective non-perceptual can illustrate the design complexity. In Table 6.4 an

Assessment	Pre check 1	Pre check 2	IQA	Post check
Rate-control	PAS	Corner-Side Edge	ppwAPC	(N/A)
Mode decision	PAS	(N/A)	SASD	Edge Pattern
Prediction	ACCS	Corner Edge	APC	(N/A)

Table 6.4 Overall proposed hybrid STDM-IQA workflow per front-end encoder stage

overview of the changes are presented for each stage of assessment, along with the type of pre-checks, IQA and the post check applied. For rate-control, a double pre-check is applied only with a fixed 8x8 sub-block IQA of ppwAPC. For mode decision, both a pre-check and post check are applied, the post check is designed to ensure that the perceptual cost is significantly large, otherwise it is not added to the STDM cost. While for prediction, like in the proposed rate-control, it has a double pre-check and where no post check is applied. In all, the pixels that have been sampled during pre-checks must be perceptually assessed with their respective algorithm. To ensure complexity remains low, the algorithm uses look up tables (LUTs) as a means to store pre-calculated values. These LUTs will contain the pixel based IQA cost in a 2D array, where original and reconstructed values are used to obtain the IQA cost. The proposed perceptual tests and assessments are undertaken with no multiplies or divides. In addition, early termination of the IQA path is possible via the use of pre-checks to ensure a low complexity design. These proposed hybrid STDM-IQA workflows rely on thresholds being met to continue along the IQA path.

6.5.2 Mode decision: SASD Complexity

The proposed modifications to mode decision involves the most number of stages and the most extensive use of edge detection. These stages have their respective block specific complexity costs and these have been summarised in Table 6.5. The table shows how perceptual complexity scales with block size, in particularly the edge detection test and SASD assessment. Under mode decision, the 2x2 edge detection is applied across the sub-block, which accounts for the large amount of operations related to edge detection. As mode decision is less intense than

Mode decision (Relative overhead)	Blockside test Pre-check 1 (Mandatory)	Edge detect Pre-check 2	SSE (Mandatory)	SASD IQA
Mult/Div			Blocksize	
Add/Sub	3	1/4 Blocksize	Blocksize	Blocksizex2
LUT	(Blocksize-2)x4	3/4 Blocksize		Blocksize
Shift		1/4 Blocksize		Blocksizex2
Compare	1	1/4 Blocksize		1
Abs	3			

Table 6.5 Mode decision: relative additional complexity overhead per square block size for IQA

prediction this additional complexity is acceptable and provides robustness. The post-check of SASD score is a comparison to avoid spurious scores being added. This is especially important as this is the final assessment stage and as discussed earlier in Section 6.3.3 the SASD cost should be applied where SSE is significant.

6.5.3 Prediction: APC Complexity

Prediction is the most varied and intensive stage of the three front-end encoding stages. Thus, the proposed perceptual cost has been designed to have very limited complexity. Unlike mode decision, which scales with block size, the complexity under prediction is largely fixed except for where block sizes can be divided by 8x8 to apply ACCS internally. All these additional processing can be summarised in Table 6.6, where for the ACCS outer and edge detect is fixed with small values. While for block size greater than 8x8, ACCS is repeated every 8x8 following the out corner ACCS on the sub-block.

Prediction: Fixed overhead			LUT	Plus/Minus	Abs	Shift	Compare
Pre-check 1	ACCS	Outer	4	3	1		
	(Every 8x8)	Inter	4	3	1		
Pre-check 2	Edge detect	Corners	12	3		3	3

Table 6.6 Prediction: additional complexity to test whether IQA should be undertaken

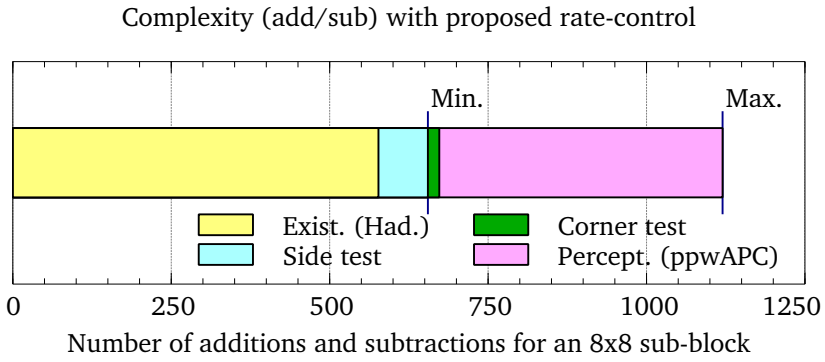


Figure 6.7 Proposed rate-control complexity of Hadamard (Had.), respective conditional tests and perceptual IQA (ppwAPC)

6.5.4 Rate-control: positive pair weighted APC Complexity

Rate-control is the least frequently called operation among the three workflows. It consists of a fixed 8x8 sub-block where additions and subtractions are used to calculate Hadamard. The proposed IQA of ppwAPC also uses additions and subtraction. While ppwAPC is complex as the 8x8 Hadamard transform, the perceptual significance tests must be applied with the minimum overhead. In Figure 6.7, Hadamard and ppwAPC take similar proportion of complexity and the double pre-check takes small fraction. This means Hadamard and the pre-checks are called continually, with the perceptual assessment of ppwAPC applied when conditions are met. As the pre-checks are minor, they will only add a fraction to the overhead costs.

6.6 Visual VCL tool

In order to simulate the respective hybrid STDM-IQA workflows, the visual VCL tool was created. This has two functions, one to simulate where the proposed IQA workflow paths would be triggered, and two to superimpose meta information (signalling bits, residue bits and QP) from the encoded bitstream on to the video frame as an image. This means that the first part is to demonstrate how the proposed hybrid STDM-IQA framework operate as a series of workflows designed for the respective front-end stages. The second part will be used later to evaluate whether the implementation results in changes to the VCL structural and meta information, resulting in bit-redistribution. Finally, by combining it with the original video sequence, assessments, including the proposed IQA and STDM-IQA workflows can be visualised.

6.6.1 Visualising assessments

Existing tools for examining the video image quality cannot show the VCL as the reconstructed video has no signalling information is available. This means that the encoded bitstream must be analysed. Existing tools which allow the bitstream to be visually analysed are commercial and/or have closed source code. The need for open source code allows simulation of existing assessments and development of new IQA methods based on real video data. Designing a tool within the video coding environment than in a mathematical software allows the development to

be orientated around future implementation within the encoder. This means that in terms of the IQA development modifying the decoder acts as a less complex prototyping environment. Since the decoder does not affect the encoding, it can also be used with other compatible bitstream files, allowing comparisons between original and proposed encoder to occur. Also, having the modified decoder as an intermediary step, encourages robust design of the framework. This is because the encoder structure has various implementations of the same assessment for different block widths. Thus, design flaws can be identified during the decoder simulation and subsequent issues within the encoder can be attributed to implementation errors.

The decoder was modified to allow the capturing, modifying and storing of relevant decoded information. This involved creating additional streams of the reconstructed frames, which would later be modified to graphically superimpose the partition grid, quantisation or bit usage. Also, the original video sources were made available so that the decoder could perform assessment between original and reconstructed frames. Finally, the frames were rendered and saved as portable network graphics (PNG) format. As described in Section 6.3.4, to provide robustness, the thresholds were refined based upon a different a video sequence, BasketballDrive. Overall, the visual VCL tool simulation assisted in refining threshold values that had been established upon observations modelled in R.

6.7 Methodology for modelling and simulating the proposed framework

In order to test the effectiveness of the respective workflows, the respective workflows were modelled in R and simulated in the visual VCL tool. The process involved to determine thresholds was a three stage process. First observations were gathered, then these were sampled from which the IQA and framework could be modelled and finally they were simulated within proposed visual VCL tool. The development of the thresholds had initially been based upon observation gathered from RaceHorses, then CrowdRun. Later, using the proposed visual VCL tool a single threshold for a sub-block size of 8x8 was set. From this threshold other values

would be scaled to form the threshold based look up tables (LUTs). To ensure robustness the visual VCL tool used pre-encoded video sequence of BasketballDrive at different bit-rates of 1, 4 and 16 Mbps to represent different levels of distortion.

6.7.1 Capturing observations for modelling

Using V16.4 of JCT-VC HEVC code base, the Hadamard 8x8 was modified to log each candidate's raw pixel value along with the Hadamard score. The SSIM was calculated based upon the code from the JM H.264/AVC codebase (Sühring, n.d.). This allowed SSIM and the proposed perceptual distortion assessments of APC and SASD to be modelled on real captured data using R (R Core Team, 2014).

Initially, the data gathered was based on standard definition video source of RaceHorses and then with a HD video sequence of CrowdRun. Prediction stage provides a variety of combinations, far greater than rate-control or mode decision. Observations were recorded using the first three frames, for RaceHorses over 1 million observations were captured of which 330,000 were unique. Later when the HD video sequence of Crowdrun was used, this produced 48 million results, of which 8.7 million were unique. Gathering and managing this large set of observations presented its own problems. As a result, a limited the number of fields were chosen per candidate assessment during prediction and this CrowdRun was set encode at a fixed bit-rate of 1Mbps. Given the technical obstacles, only a single bit-rate was used. The choice of a low bit-rate for a highly active HD video sequence of CrowdRun was considered an extreme example, however, at that bit-rate it is highly likely to contain perceptually significant distortion.

Using R, a set of 30,000 randomly sampled observations were selected, where the random number generator seed was set to 42, which ensures reproducible results using the same original observations. Given this sampled dataset, the IQA applied to all and thresholds for the framework was initially set by examining the summary statistics of STDM and 1-SSIM distribution as thresholds were adjusted. Initially, there would be a pre-design based threshold, then the dataset median was considered and finally a binary friendly version was explored where the 1-SSIM would capture perceptually annoying observations.

With SASD and APC values, the observations were perceptually filtered according to their respective workflows and then assessed against SSIM. For rate-control

where distortion activity is used, the first frame of CrowdRun a HD resolution (1920x1080 pixels) video sequence was used. This produced 32,400 observations, (1920x1080)/(8x8). For each 8x8 raw luma array, an averaged value was gathered and this was used to create a JND profile of the frame This allowed the ppwAPC perceptual filtered frame to be measured against the reference JND profile of the frame.

6.7.2 Visual simulation via the visual VCL tool

Following on from the modelling stage, was the visual simulation stage, using the visual VCL tool. The visual VCL tool was designed to perform refinement of the initial modelling stage based thresholds, to ensure that the settings chosen to reflect perceptually significant regions. For this another HD video sequence was used BasketballDrive, to avoid making a very specific solution. However, in the results, the visual VCL tool applies the final thresholds on CrowdRun which was not used during development by the visual VCL tool. Due to the limit number of HD video sequences, this approach provides a limited set of robustness, meaning the thresholds are determined by a few sets of video sequences. However, this does allow the remaining available HD video sequences to for testing the proposed implemented framework. Also, visual analysis using BasketballDrive was performed with video sequence encoded at different bit-rates of 1, 4 and 16 Mbps all encoded using the random access profile.

6.7.3 Generating heat maps

The visual VCL tool provided visual feedback in understanding where the IQA path of the proposed workflow would trigger on the given video frame. In order for the graphics output to be intuitive, heat maps were created based upon the meta and assessment information. To maximise definition between the range of scores, the colour would progress from red, the highest, to blue, the lowest, as in first five colours of the rainbow (red, orange, yellow, green, blue). In order to aid visualisation, the underlying luma information was amplified to ensure visibility of texture to aid visual analysis. Similarly, colour intensity would reflect the ratio of the meta information or assessment score. This is especially expressed for signalling information, where the four states between these colours were used as quadrants

when scaling range of scores. For example QP < 14 goes from blue to green, QP > 13 and QP < 27 starts at green goes onto yellow, while QP > 26 and QP < 40 moves from yellow to orange and finally QP > 39 to QP 51 shifts from orange to red.

While STDMs and the proposed IQAs are pixel-based, the graphical output are designed to be visually representative than being technically accurate. This is because a fixed window size of assessment were used either 8x8, 16x16, 32x32 or 64x64, based upon the configuration file than the dynamic meta information. Consequently, the configuration file contains settings by which the scaling may be applied including bit usage where bits per coding unit may be defined. In order to establish a reference when comparing each of these graphics the underlying frame image is from the original video sequence and not the reconstructed video, which assists when analysing the results.

6.7.4 Adjusting thresholds and scaling of IQA scores

Following on from this two stage process, the thresholds upon which the IQA path may trigger and the scaling which represents the magnitude of the IQA score was liable to change. For SASD the threshold remained the same, at 512, equally the downscaling was also unchanged at right shift of 7, which is 1/128. For prediction, the high level of false trigger during BasketballDrive led to a doubling of the threshold from 128 to 256, see Section 6.3.4. In rate-control, the threshold based upon the asymmetrical sides test initially designed guide of 64. Using observations from RaceHorses video sequence the threshold changed to 84 before changing again to 48, when CrowdRun, a HD video sequence observations was used.

Finally, in prediction and in rate-control which share the same norm space, the downscaling of total IQA score prior to adding to the STDM score was changed to 1/128, which matches SASD downscaling. This new figure was obtained by using rate-control observations in prediction modelling as extreme differences. While this simplifies the downscaling of IQA scores across all front stages, this means that in rate-control and prediction IQA score represent < 1.25%, than the target of $\approx$ 10%. This means that any emphasis by the use of an IQA score should be subtle during rate-control and prediction. Under rate-control this reduces the impact of

false triggers and in prediction it allows similar candidates to be selected. Please note, that the results for the modelling stage do not include the changes following adjustment.

6.8 Results for modelling and simulation

The results presented here are IQA on video frame, threshold modelling, ratio of IQA over STDM and visual simulation of proposed STDM-IQA workflows respectively. The modelling of the proposed hybrid STDM-IQA for each respective workflow is based upon captured raw data. This provides an overview for the likelihood that a IQA related score will be applied as an addition cost to existing STDM score, and whether the IQA cost within $\approx$ 10% of the respective STDM score. However, it is important to understand where these proposed IQAs may be applied, for that the visual VCL tool is used. The proposed visual VCL tool applies the proposed IQA and as part of the respective hybrid STDM-IQA workflow on a set video frame, to illustrate which regions in the frame triggered by the IQA path.

6.8.1 Modelling the hybrid STDM-IQA framework

The proposed hybrid STDM-IQA framework, has been defined as a series of tests and an IQA for each of the front-end encoding stages. These solutions are simulated to illustrate their effectiveness at being able to target perceptually significant distortions. This means that the observations gathered must be processed to the respective proposed perceptual solution.

The results are presented as a series of density plots. Density plots are similar to histograms, however, density plots are an instantaneous representation for the distribution over an integral period, which can allow scores to be greater than 1. Under R, the bandwidths for density plots are chosen automatically, which allow a smooth representation compared histograms.

Figure 6.8 is the JND profile emulating the proposed perceptual rate-control activity assessment. While Figures 6.9 and 6.10 are 1-SSIM profile of proposed IQAs of SASD and APC respectively.

JND profile vs proposed rate-control

The JND profile shown in Figure 6.8 is based upon the averaged luma value of each 8x8 array. This explains why both the frame JND and filtered profile

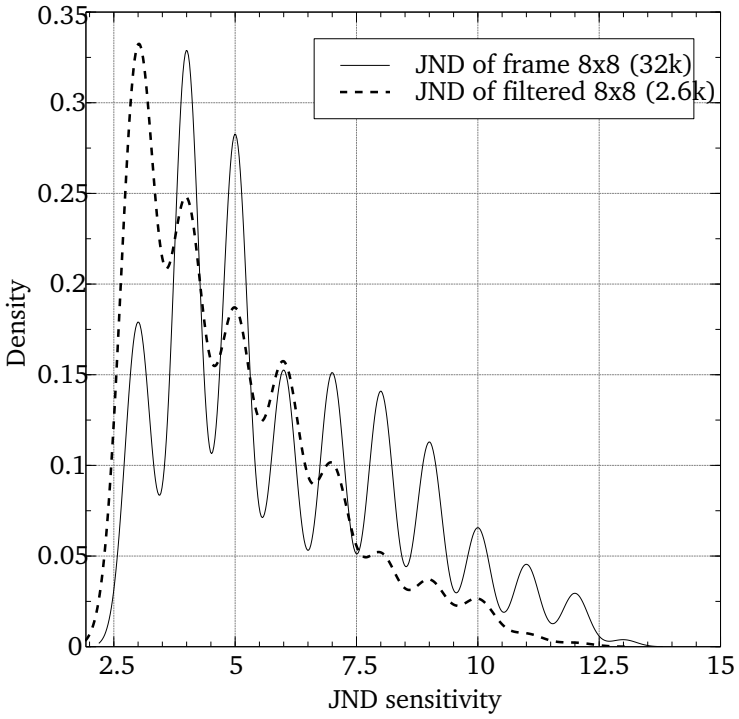


Figure 6.8 Density plots of JND on frame, unfiltered vs proposed double threshold.

oscillates. From the JND profile of the frame the highest sensitivity is localised around the lower to middle end of the JND sensitivity scale. This is reflective of the content which is set outside during the day. The filtered JND profile based on the proposed perceptual rate-control workflow has a high density of extremely low JND sensitivity observations. Also, the density for middle to higher JND sensitivity observations are significantly less. From a total of 32.4k, $\approx$ 8%, 2.6k observation underwent the proposed perceptual rate-control activity workflow. This means it will more likely to affect those blocks which have low JND sensitivity threshold (brighter areas) than those with high JND sensitivity threshold (darker regions).

1-SSIM profile vs proposed mode decision and prediction

Figures 6.9 and 6.10 both show the reference 1-SSIM profile of the observed data as twin peaked around zero and one. The subsequent modelled workflow of SASD and APC are largely centred around the peak along 1-SSIM which is just before 1. Under 1-SSIM, zero represents perceptually indistinguishable, while one means it is considered perceptually poor. The observations were gathered from prediction than mode decision, this means the variation in obserations is greater than mode decision would typically have. However, the SASD profile under

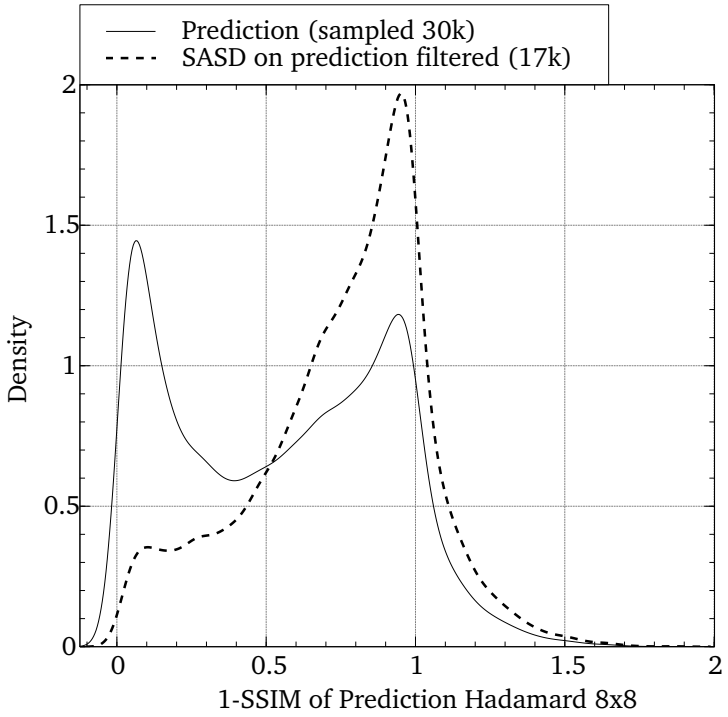


Figure 6.9 Assessing block perimeter using SASD: Density plots of proposed perceptual filter vs same filtering based on metric differences and unfiltered data.

prediction observations in Figure 6.9 does have the desired effect. It is worth nothing that over half the observations, 17k out of 30k, were accepted resulting in the broad base for the filtered profile. In the proposed IQA workflow for prediction in Figure 6.10, the base is narrow and the peak is double that of SASD, yet the number of observations are 5%, 1.5k out of 30k, $1/10^{th}$ of SASD. As prediction is the initial stage, this narrow focus is acceptable to regulate IQA where relevant especially due to the volume of candidates and the variety of sub-blocks sizes. In all, SASD in mode decision has a broad acceptance of what is classed as perceptually significant distortion, while prediction changes are restricted to where 1-SSIM is high.

6.8.2 Ratio of filtered IQA score vs. respective STDM score from modelled framework

Continuing on from with the filtered IQA scores, these can examined by their ratio of IQA against STDM, similar to the heat maps shown in Figures 5.5 and 5.8. As stated when the IQA only heat maps were produced, downscaling or capping

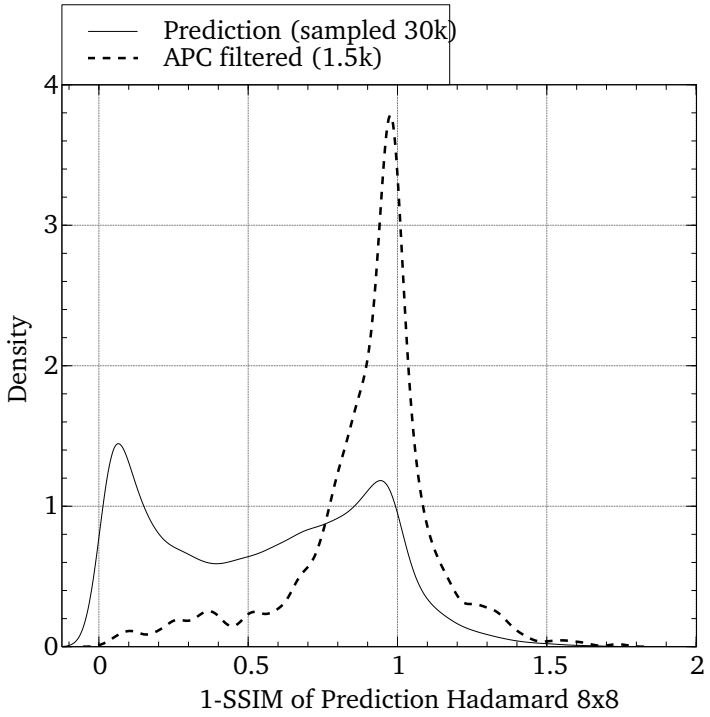


Figure 6.10 Assessing block perimeter using APC: Density plots of proposed perceptual filter vs same filtering based on metric differences and unfiltered data.

would be potentially required. The downscaling and cap have nominally set at $\approx 10\%$ of the STDM scores. Therefore, the intent of showing the ratio of the filtered IQA scores over the STDM scores is understand whether they are satisfying this target.

In the modelling the scores for ppwAPC in rate control and APC in prediction by are downscaled by 1/32 and 1/16 respectively, while for SASD (aimed at mode decision), the scores are doubled. In Figure 6.11 virtually all the observations can be accounted for within the target of $\approx 10\%$. This means that clipping is not required, especially, as the rate of decline is dramatic at the 5% mark. For SASD, the filtered results in Figure 6.12 illustrate the spread exceeds $\approx$ 10% mark, meaning that clipping is required at 12.5%, which is equivalent to $1/8 \cdot SSE$ or by right shift SSE by three. Finally, prediction APC scores in Figure 6.13 shows a broad spread of distribution of values around the 5% mark. These scores slightly exceeds the ratio of $\approx 10\%$, meaning no clipping is required, as any additional processing would add to the complexity overhead. Overall, the downscaling of perceptual scores can be

shown to approximately be within or around the 10% maximum, however how this performs visually needs to be examined.

6.8.3 Visual VCL Tool

The visual VCL tool is designed to show the breakdown of the signalling choices including the frame partitioning into the respective sub-blocks. The structure of these partitions is influenced by the respective assessments scores in the stages of rate-control, prediction and mode decision. Through the visual VCL tool, it is possible to simulate assessments, signalling information and the proposed workflows. The results shown in Figures 6.14 to 6.17 include bit distribution, and forms of assessment or activity. While the results presented here in Figures 6.18 to 6.21, are of the proposed framework, illustrating where the respective pixel-based IQA would trigger.

Visualising existing STDMs and IQAs

The results in Figures 6.14 and 6.15 illustrate the decoded output graphics of theRracehorses video sequence that was encoded at 128k, 256k and 512k. In Figure 6.14 the blocks which contain residue information are shown, with different

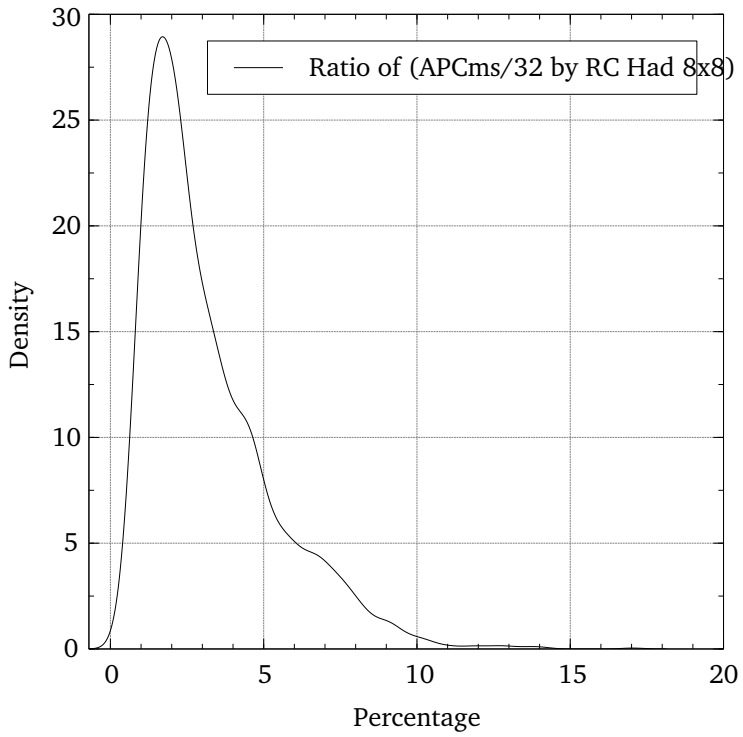


Figure 6.11 Density plot: Ratio of APCms by RC Had on filtered data

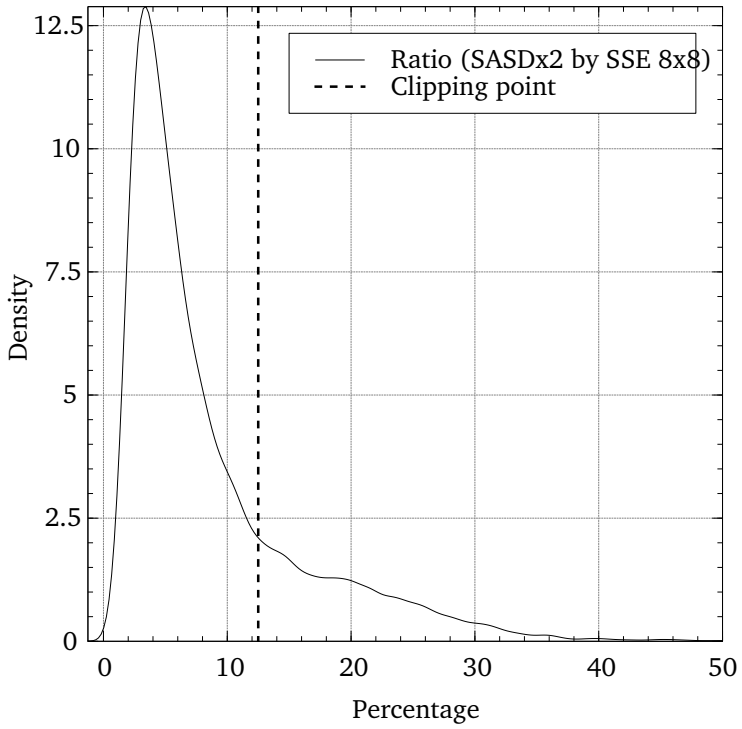


Figure 6.12 Density plot: Ratio of SASD by SSE 8x8 on filtered data

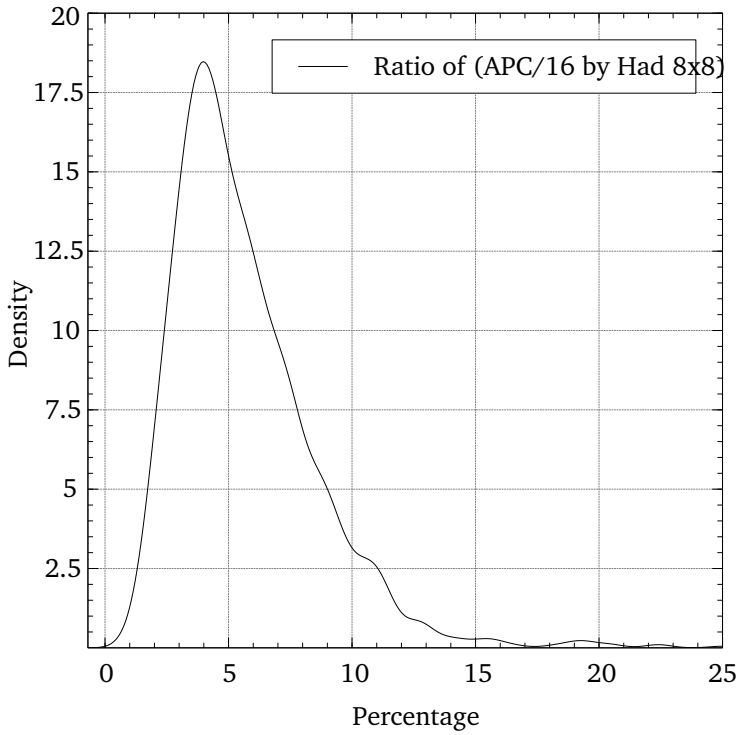
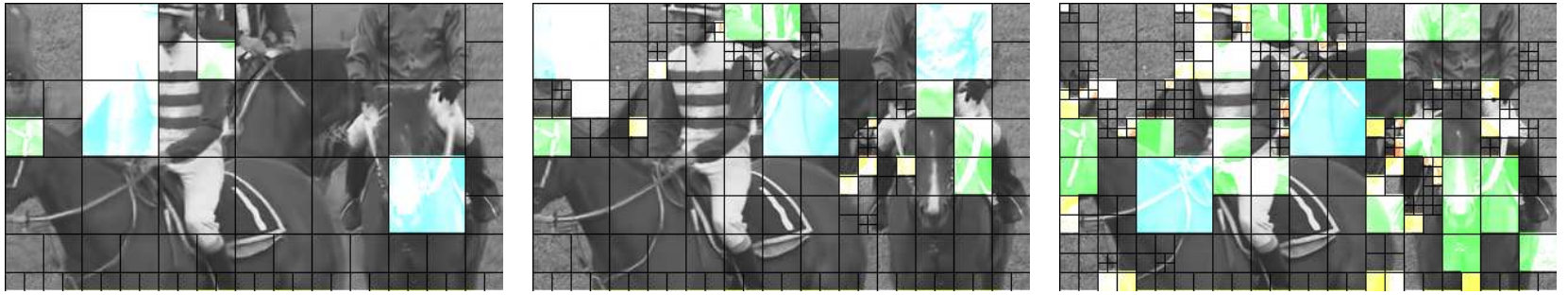


Figure 6.13 Density plot: Ratio of SASD by SSE 8x8 on filtered data

block sizes highlighted by their respective colour. In Figure 6.15 the bit usage of each LCU is highlighted by colour based upon the given setting, in this case maximum value is set to 128 bits per LCU. While in Figure 6.16 existing distortion metrics are applied on the video sequence Racehorses frame 7 at a bit rate of 256k. For Figure 6.17, the activity of the given original frame is shown using rate-control Hadamard and JND, both are designed to operate on the original frame.

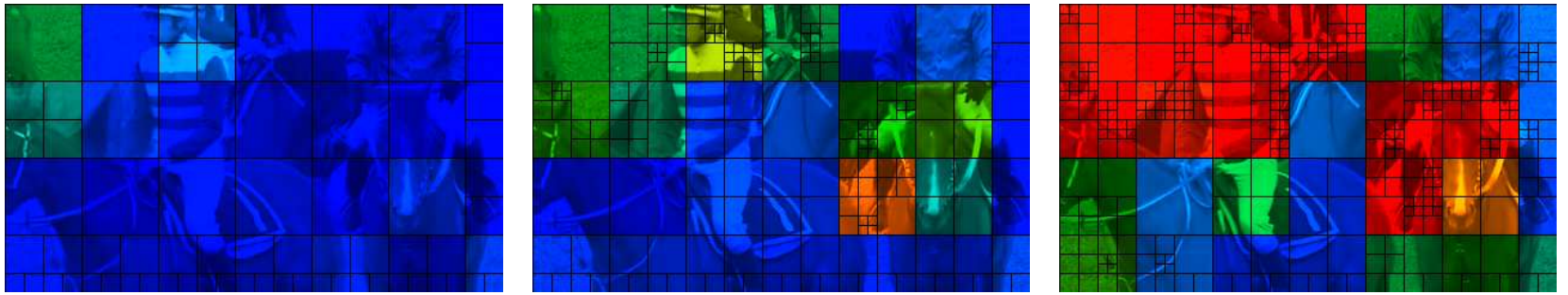
Visualising proposed IQA workflows

Visualisations of the proposed IQAs are shown in Figures 6.18 to 6.21, this is based upon the decoded CrowdRun video frame 71. The decoded frame is from an encoding were the first 150 frames were encoded using HEVC V16.6 at 4Mbps. Each of the frames have been cropped, with three regions highlighted by a white rectangle. These regions denote different density of activity, the upper box is where the sky is least dense, the middle box where the crowd has high density of textures and the bottom of runners is where medium density of activity. While the coloured markers represent where an observation has undertaken the proposed IQA workflows path on top of the respective existing assessments.



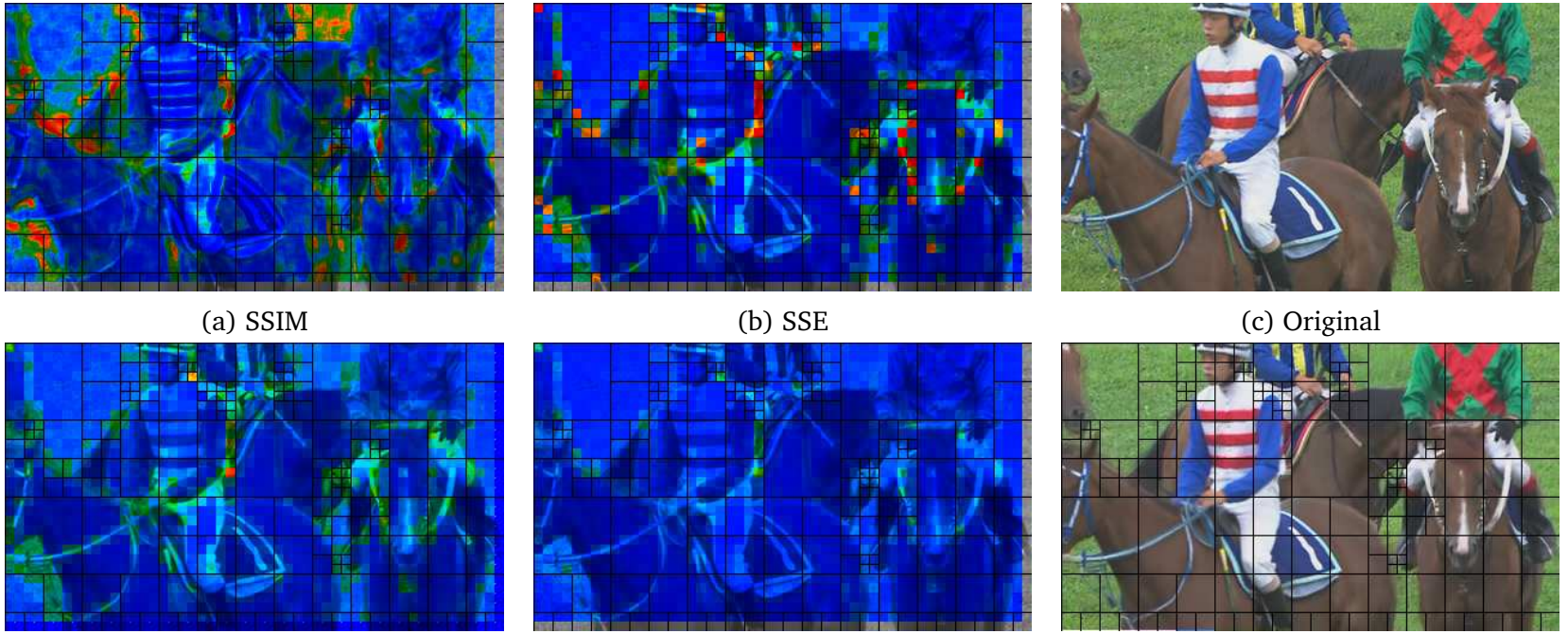
(a) Racehorses 128k (b) Racehorses 256k (c) Racehorses 512k

Figure 6.14 Video sequence racehorses frame 7, partitioned with highlighted blocks, where blue is 64x64, green is 32x32, yellow is 16x16, orange is 8x8 and red is 4x4



(a) Racehorses 128k (b) Racehorses 256k (c) Racehorses 512k

Figure 6.15 Video sequence racehorses frame 7, bit usage per coding unit with partitioning, where blue is low, red is high, heat map largest coding unit (LCU) maximum 128 bits



(a) SSIM (b) SSE (c) Original

(d) Hadamard (e) SAD (f) Reconstructed

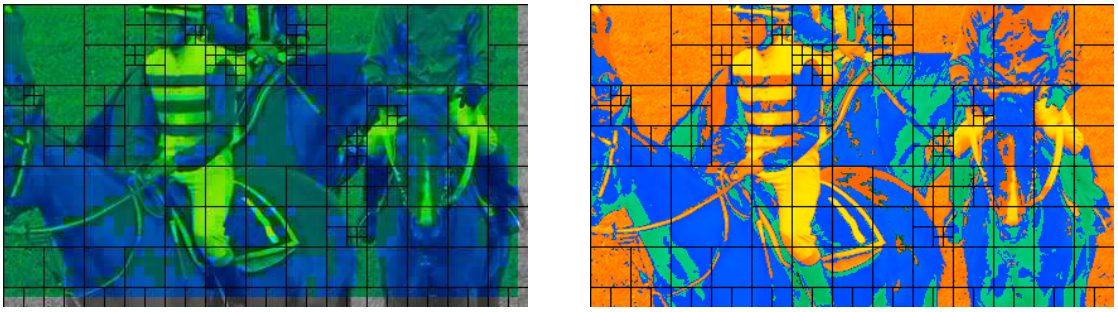
Figure 6.16 Video sequence racehorses frame 7, distortion assessment with partitioning, where blue is low, red is high, encoded in HEVC with target bitrate of 256kbps

6.9 Discussion of results

The results highlighted different aspects of the proposed hybrid STDM-IQA framework. Here the respective proposed IQA workflow paths were evaluated by filtering recorded observations and visualising them on the proposed simulation tool. The filtering and visualisation indicate that the proposed IQA workflows are sensitive to activity and distortion in brighter regions. The threshold for mode decision and prediction may need to be revisited as the number of recorded observations from the data modelling and visual assessment from the simulation could be of concern. However, as the mode decision and prediction is applied to a variety of block sizes, adjusting this could potentially affect the overall complexity for encoding a video sequence. Overall, neither the modelling nor simulation is a complete representation of the entire front-end encoding process. This is because the modelling is based upon prediction candidates of 8x8, while the simulation is based upon the bitstream, where only the best mode decision candidate is available.

6.9.1 Modelling the hybrid STDM-IQA workflows

The respective hybrid STDM-IQA workflows were implemented inside R where recorded observations were used to measure their responses. Overall the resulting reduction of observations for rate-control, mode decision and prediction workflows represented 8%, 56% and 5% respectively. For rate-control, the filtering against JND sensitivity is low for high luma values as Figure 6.8 which means additional bits would be allocated to those brighter sub-blocks. For mode decision in Figure 6.9, despite the high number of candidates it is likely that an IQA score of SASD is



(a) Rate-control Hadamard (b) JND

Figure 6.17 Video sequence racehorses frame 7, activity, rate-control Hadamard and JND, where blue is low, red is high

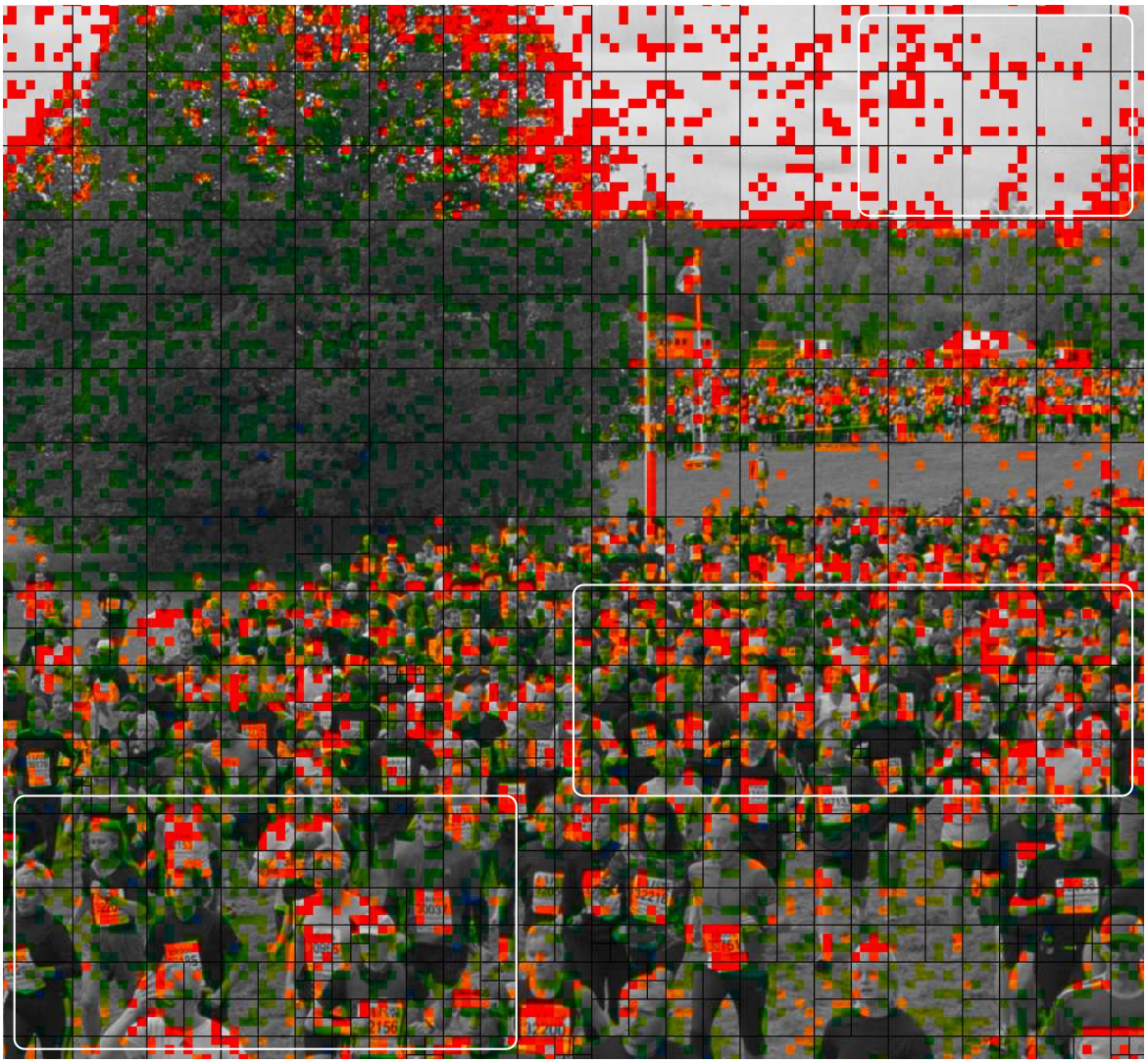


Figure 6.18 Simulated: Rate-control - Hadamard 8x8 with msAPC, highlighted were affected

applied where distortion is perceptually annoying. While the very narrow selection in Figure 6.10 demonstrated that the pre-checks are effective at filtering false positives. This is especially useful, as in the IQA on video frame result of Figure 5.9 highlighted the risk of APC being applied everywhere. In all the thresholds are strict for the proposed ℓ_1 IQA and loose for the proposed ℓ_2 IQA. As the volume of candidates are high for the prediction stage and APC is highly dynamic, this restriction is acceptable as prediction is complexity sensitive.

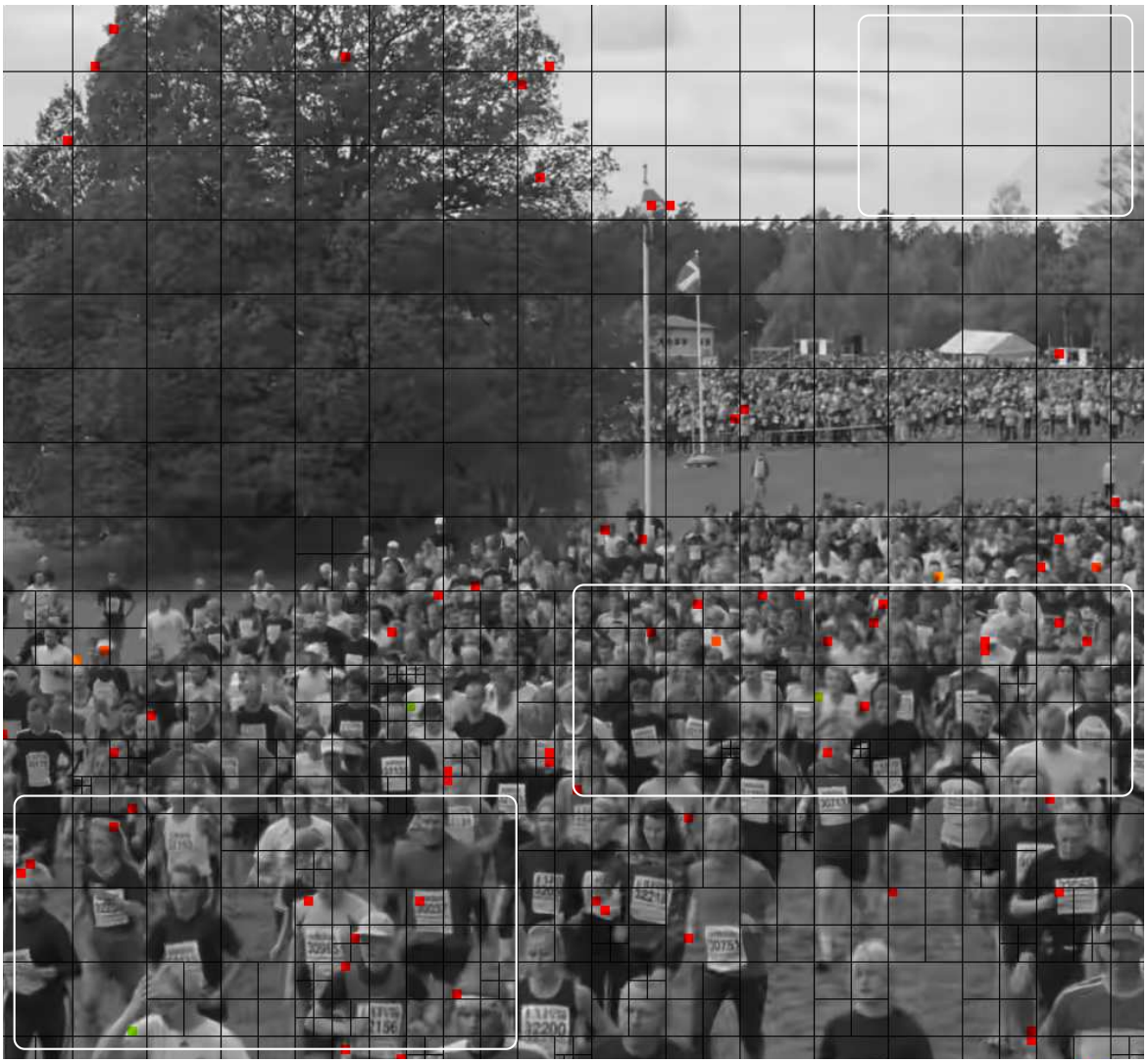


Figure 6.19 Simulated: Mode decision - SSE 8x8 with SASD, highlighted were affected

6.9.2 Compatibility of IQA scores

During the literature review and previous chapter, SSIM was described as where scores did not satisfy the $\triangleq$. In the proposed IQAs, the $\triangleq$ was not discussed as the proposed IQAs were designed to be an additional cost to STDMs, limited to $\approx$ 10% of the STDM score. The choice of $\approx$ 10% was deemed a suitable maximum as a means to discourage a candidate over another. A larger percentage IQA cost would risk eliminating candidates, than re-order them and potentially allow another overlooked yet perceptually annoying candidate to be selected. The choice of maximum of $\approx$ 10% may not be the optimal value. This is especially important

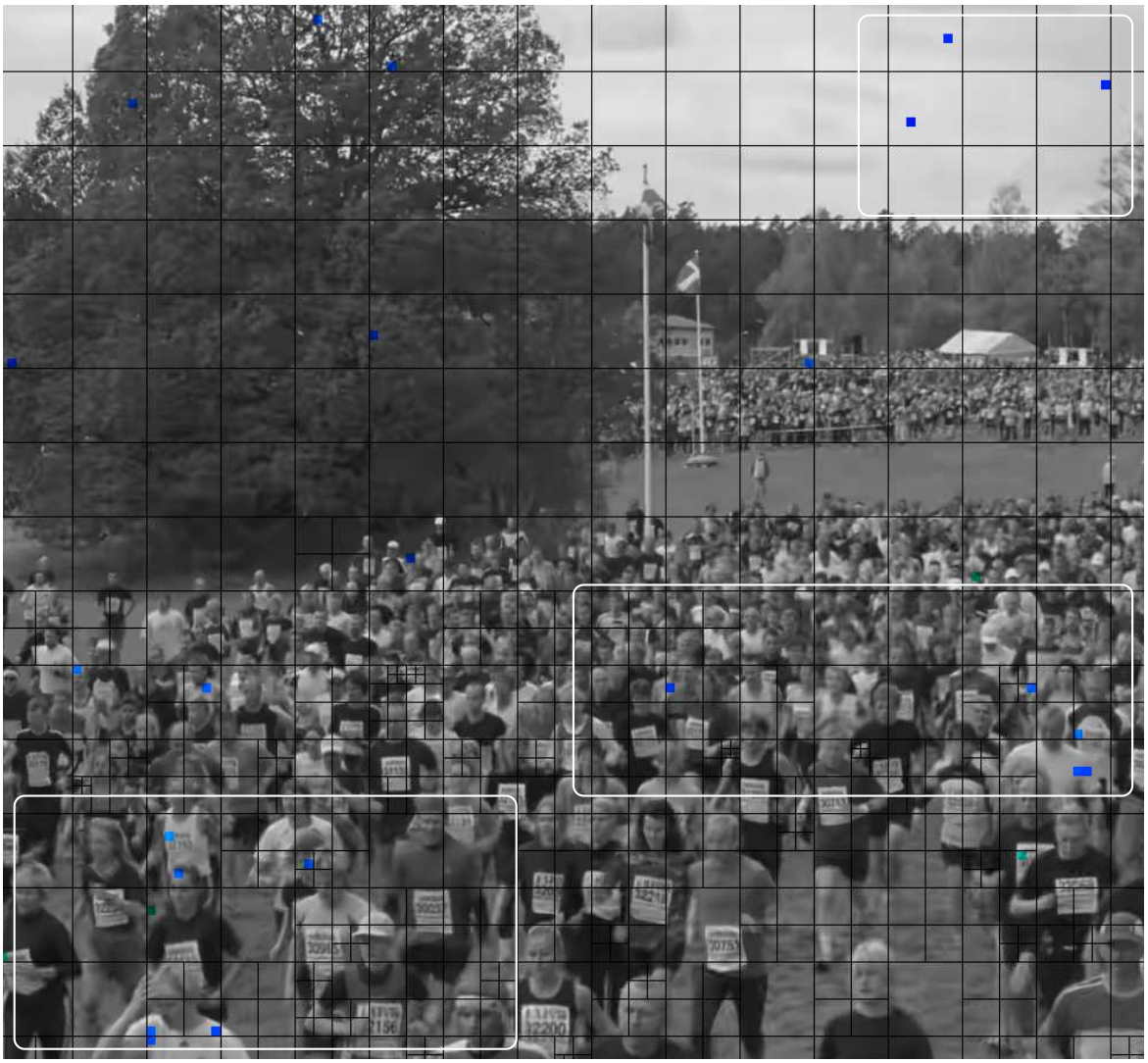


Figure 6.20 Simulated: Prediction - SAD 8x8 with APC, highlighted were affected

when considering the SSIM-STDM shared space that was discovered in the previous chapter, which led to the proposed R-D equation where κ was introduced to skew distortion scores. As the shared space reflects where the $\triangleq$ could be met, an IQA cost is limited by how much it can stretch the STDM score. The filtered observations during modelling were analysed for their ratio of IQA score over the respective STDM score. In all, these density graphs were shown to meet the target projection of $\approx 10\%$ or could be capped with minimal cost.

This $\approx$ 10% shown to be possible during modelling stage of the framework, however, later the same data used for rate-control was considered for prediction as an extreme differences. Under these circumstances, the prediction was downscaled

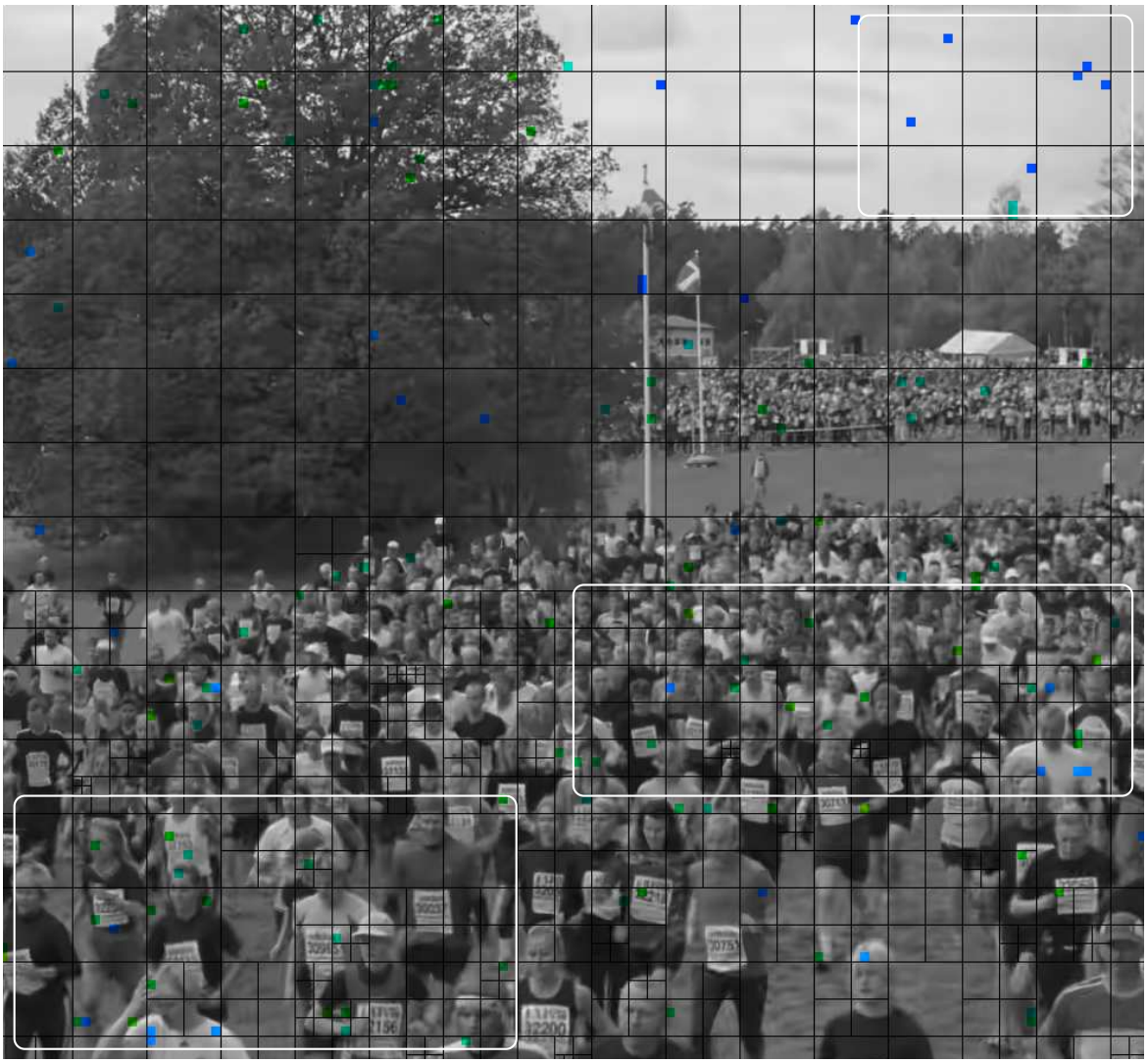


Figure 6.21 Simulated: Prediction - Hadamard 8x8 with APC, highlighted were affected

to 1/128, effectively, to <1.25 % of the STDM (SAD or SATD). This does means a simpler design for the implementation stage, however, it does leave the question open to what is a suitable level for κ? To answer this perhaps more observations based upon video sequences at different bit-rates. Unfortunately, both access to HD video sequences are limited and handling observations can be difficult. This question will remain open, however, for now the downscaling used in ppwAPC (rate-control) and APC (prediction) will be 1/128.

6.9.3 Modified decoder

Figure 6.14 reflect those sub-blocks which have residue information, where the colour of highlighted blocks are lighter as quantisation is higher. This is shown, albeit faintly, on the left hand side of the frame, with the horse's eye, where the same sized sub-block is used yet the intensity varies. Within these figures, as the bit rate increase, the number of sub-blocks with residue increases faster than the decrease in quantisation. For where bandwidth is less, particularly during RDO, combinations of larger sub-blocks are sought. In contrast, Figure 6.15 illustrates how the spread of bit usages moves from mostly uniform to non-uniform per LCU. This is due to the higher bit budget, which allows for increases in partitioning for tracking the movement via motion vector signalling. In all, these partitioning choices are areas where assessment can determine the distribution of bits and choice of signalling.

Also, the four assessments of SSE, SSIM, SAD and SATD have been shown graphically in Figure 6.16 for the same video sequence frame. Together they highlight that SSIM is complex, and visually SSIM may not react to distortion in the same manner as STDMs. Since assessment affects the choice of partitions, it is important that an IQA is able to preserve the inherent perceptual clues. The activity maps shown in Figure 6.17 are seen to be quite different. The first shows existing rate-control, based upon Hadamard, Figure 6.17a, where the given 8x8 image is assessed as if it were the differences. While the second, Figure 6.17b applies the perceptual model of JND expressing the HVS on the frame. The Hadamard activity is flat compared to the dynamic changes in the JND frame. This indicates that perceptual activity presents the opportunity to ensure that bit distribution for regions of perceptual interest.

6.9.4 Visual simulated proposed algorithms

For rate-control, the simulation is applied on the original video source and not on the reconstructed frame. This means that an 8x8 pixel array gathered can be largely similar to how it will be processed when implemented in the encoder. In Figure 6.18, the modified decoder shows activity as detected across the frame, largely where there are changes in lighting. This is shown particularly where the

trees and sky meet, or the cloud texture and shirt numbers for the runners. In the reference frame, rate-control (without proposed) was uniformly red, indicating high activity throughout. This suggests that further additional cost would be added to areas marked as perceptually significant, which potentially results in a non-uniform distribution of bits per LCU.

In the encoder, mode decision evaluates different combinations of block sizes using SSE to find the lowest cost. Under the visual VCL tool, a much simpler version is expressed, one where a fixed 8x8 window of SSE is applied along with perceptual checks before considering SASD. The resulting decoded frame of Figure 6.19 illustrates with markers where perceptual processing would occur. This is far fewer than rate-control with ppwAPC. The perceptual dominant side test threshold for SASD is significantly lower than rate-control ppwAPC, 8 and 48 respectively, however, coverage of edge detection is significantly higher. This means that the likelihood of SASD with SSE occurring is significantly less for the same 8x8 block size. This trigger rate of SASD would be similar for other block sizes if linear threshold is used and potentially far less for larger blocks under non-linear. The markers representing where SASD would be applied correspond to high changes in textures, indicating that the edge detection across the sub-block was high. As edge detection is based upon the perceptual SASD score, these occur on bright high intensity areas that boundary darker regions.

For prediction simulation, Figures 6.20 and 6.21 cover SAD and Hadamard respectively with APC, both apply the same ACCS and corner edge test, yet have each has a different response. They tend to trigger on the same areas, with Hadamard version having additional markers elsewhere. The proposed perceptual distortion difference is particularly sensitive to where original luma values are high. This is shown by the location of the markers, which refer to where the original high luma values may be substantially different due to averaging with a neighbouring darker region. The number of markers shown are very low, similar to the simulation where 5% was observed, however, during prediction the variety of block sizes and volume of candidate choices are high. Consequently, the visual VCL tool highlights that having a higher trigger rate (lower test threshold) for applying IQA score does risk increasing the encoding complexity.

Overall, these assessments do exhibit sensitivity to lighting changes, with rate-control based ppwAPC showing the greatest level of triggering. While for mode decision and prediction the perceptual assessment path occurs less frequently. This means that as quantisation changes those candidates which undergo perceptual IQA could influence the VCL structure and the final encoded video bitstream.

6.10 Summary of chapter

From the previous chapter the proposed IQAs were shown to be perceptually aware yet highly active, to stabilise them, a framework was proposed. The STDM-IQA framework was devised to allow IQAs to be called based upon meeting conditions. This meant developing tests and establishing thresholds to produce this framework for each encoder front-end stage of prediction, mode decision and rate-control. To evaluate the proposed framework, each stage was modelled in R using observations gathered from video sequences. Furthermore, the proposed visual VCL tool was produced to aid development in simulating and also shown that it would useful to analyse the VCL when the proposed framework is implemented. While the IQA scores were for rate-control and prediction were downscaled further than the initial $\approx 10\%$ figure, it could be open for further investigation. Finally, it has been established that the framework is able to isolate perceptually annoying distortion and triggering on perceptually significant regions. This will mean that the framework will be implemented into the High Efficiency Video Coding (HEVC) encoder to enable a low complexity sub-block level PVC solution.

Chapter 7

Implementation and test set-up for the hybrid STDM-IQA framework

From the previous chapter, a proposed hybrid STDM-IQA framework has been designed to use STDMs with the proposed pixel-based IQAs, subject to condition being met. This design would allow regulating complexity to make PVC viable at the sub-block level and in this chapter be applied to a hybrid block-based encoder of HEVC. Inevitability, technical challenges will be encountered during the implementation to adapt the framework for HEVC and this will be discussed in this chapter. Furthermore, a test set-up and strategy is required for the subsequent stage of testing, this will also be presented, based upon evaluating for low complexity PVC.

7.1 Technical challenges for implementing the proposed PVC workflows

Implementing the proposed PVC workflows in HEVC requires understanding how the hybrid block-based encoder structure depends upon STDMs for assess-

ment. This dependence on STDM means that only the differences are passed during assessment, while for PVC the original pixel values are required. This issue has been faced with H.264 in the initial set of experiments, however, under HEVC the technical procedure is different, meaning that additional work is required. An example which requires its own HEVC related solution is that of motion estimation for b-frames, when differences are calculated, the eventual reconstructed pixel can be beyond the pixel bit depth range, which risks the stability of any PVC solution.

7.1.1 Motion estimation modification for b-frames to support PVC

The motion estimation used in bi-prediction assumes that the source and destination pixel values have minor variation. Bi-prediction can refer to co-located sub-blocks across multiple frames which during random access where hierarchical encoding is encouraged (JCT-VC, 2016; Sullivan et al., 2012). Therefore, the likelihood of displacement is higher during random access. The difference for motion estimation in bi-prediction is shown in Equation (7.1).

$$Diff = 2 \cdot Dst - Src \tag{7.1}$$

where $Diff$ is difference, dst is destination and src is source. This can produce both negative and out of range values as shown in Table 7.1. Under STDMs, these difference would be resolved to a positive result. While for IQAs, the negative and out of range differences are unsuitable. For LUTs, a negative values for would be invalid and potentially destabilise the encoder.

To avoid this issue, the pixel value must be tested and clipped if necessary. The previous method of clipping had been abandoned by the software maintainers due to reasons of performance. Therefore, a proposed method which results in choosing the destination pixel value was applied (JCT-VC HEVC, 2013). Under a short five frame GOP test, for 8, 10 and 12 bit video sequence the timing differences was within $\pm$0.5%, as shown in Table 7.2.

7.2 Design of the proposed in-loop PVC solution

Underlying the proposed hybrid STDM-IQA framework are the proposed pixel-based IQAs, applied at the native sub-block level. The construction of these

		Dst								
		0	1	16	64	128	192	240	254	255
Src	0	0	2	32	128	256	384	480	508	510
	1	-1	1	31	127	255	383	479	507	509
	16	-16	-14	16	112	240	368	464	492	494
	64	-64	-62	-32	64	192	320	416	444	446
	128	-128	-126	-96	0	128	256	352	380	382
	192	-192	-190	-160	-64	64	192	288	316	318
	240	-240	-238	-208	-112	16	144	240	268	270
	254	-254	-252	-222	-126	2	130	226	254	256
	255	-255	-253	-223	-127	1	129	225	253	255

Table 7.1 Invalid pixel values for bi-prediction motion estimation.

Video	Resolution	BitDepth	Frames	Bit-Rate	Y-PSNR	Time
RaceHorses	416x240	8	30	1.04%	-0.0579	0.49%
OldTown	1920x1080	10	20	0.07%	-0.0008	-0.48%
Traffic	2560x1600	12	10	0.13%	0.001	-0.24%

Table 7.2 Clipped pixels for motion estimation bi-prediction on five frame test.

pixel-based IQAs were previously shown to operate at the native sub-block level and within their respective hybrid STDM-IQA workflows. Bringing the proposed hybrid STDM-IQA framework to HEVC means integration with existing parts of the encoder. As this is a sub-block level PVC solution, this means that the proposed framework phases of pre-check(s), IQAs and post-check will elongate existing assessment paths. However, as described in the previous chapter during the design of these workflows, the thresholds used in the perceptual significance tests of pre-check(s) and post-check act as points of termination. Subsequently, reflecting the proposed hybrid STDM-IQA framework in the design of a hybrid block based encoder, means understanding how these changes affect operation at the sub-block level. This section will illustrate how the proposed hybrid STDM-IQA framework will be implemented as part of a low complexity sub-block level PVC solution.

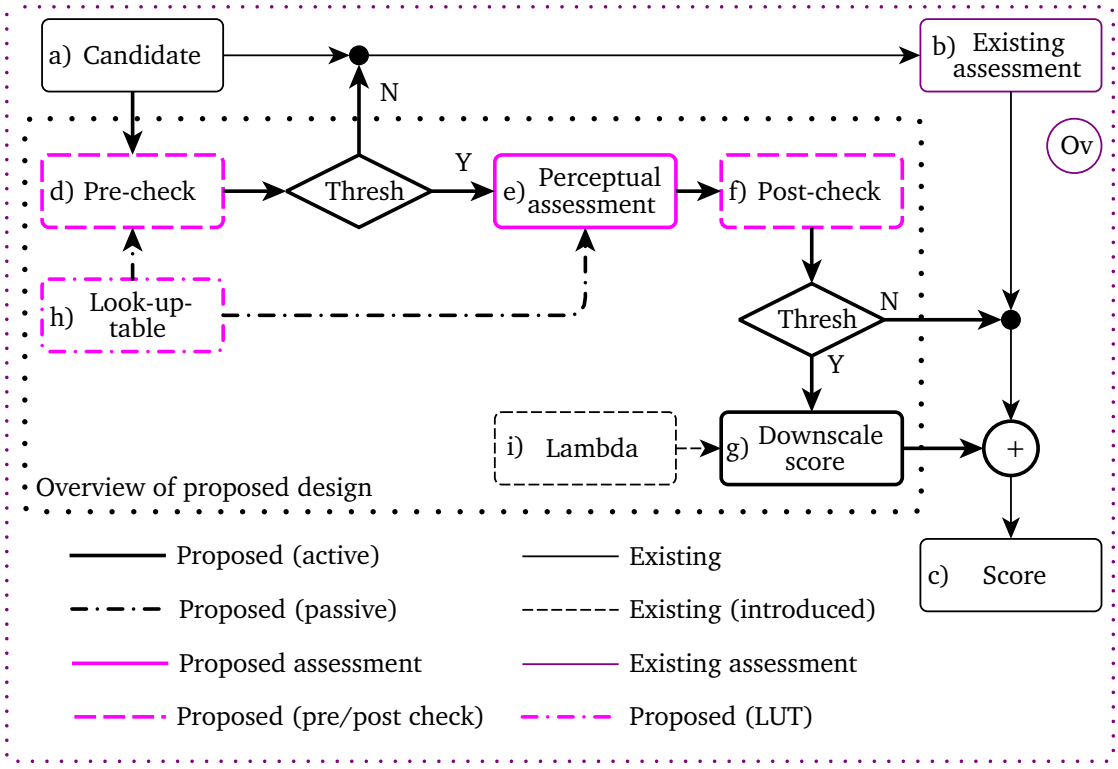


Figure 7.1 Common proposed design overview (Ov)

7.2.1 Common design approach for proposed in-loop PVC solution

From the previous chapter, a common proposed design overview, labelled as 'Ov' within the illustration in Figure 7.1, to act as a basis for the respective stages of rate-control, mode decision and prediction. This overview design can be described by its paths; where a) to c) refer to the STDM path, d) to g) as the perceptual path, while h) and i) represent the look up table (LUT) and lambda (λ). The LUT is a means to store pre-calculated values of the respective perceptual algorithms, which can reduce the encoding time. In the respective diagram, lambda, λ, is used to affect the level of perceptual cost during the final downscaling stage. However, this was an area which was not investigated during the design stage and thus will not be implemented into the proposed modified HEVC encoder. This means, that the implementation will provide a sub-block PVC solution based upon perceptual assessment and activity which will be expected to perform redistribution

of bits and signalling. Had λ been investigated then it would mean that perceptual quantisation by way of λ_p would be possible and perceptual gain could be analysed. Nonetheless, the designs presented here forms the basis of implementing a low complexity in-loop PVC.

7.2.2 Proposed perceptual rate-control activity

The proposed perceptual rate-control workflow is based upon activity of the original frame. This can have an impact on the distribution of the bits allocated to different frame types. Similar to rate-control, the perceptual activity assessment is based upon a form of Hadamard transform, if applied it doubles the complexity overhead. Therefore, the proposed rate-control design in Figure 7.2 differs from the common design in that it has a double pre-check and no post check. The double pre-check applies two different perceptual techniques, sub-block sides and corner edge detection to minimise when the additional complexity is executed. Then for those 8x8 sub-blocks deemed perceptually significant the perceptual activity assessment

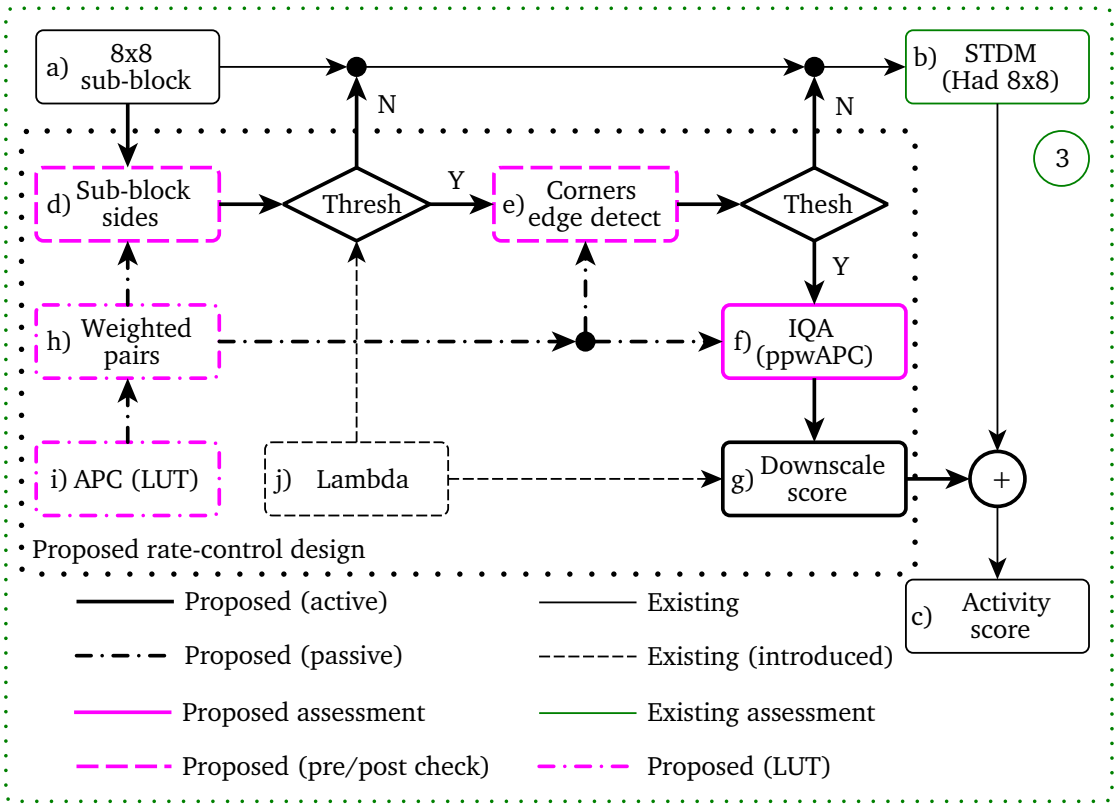


Figure 7.2 Proposed rate-control design

is applied. In perceptual activity assessment, the use of a LUT on activity must apply weighted pairs to ensure adjacent pairs of greater significance than non-adjacent pairs. This ensures sharp changes in lighting, by boundaries or textures are captured, compared to slower transitions within the sub-block. Finally, the score is downscaled by a fixed rate of 1/32, which λ could affect. Consequently, the proposed rate-control differs from the common design with the double pre-check and no post-check, apart from which it is otherwise very similar to the common design.

7.2.3 Proposed mode decision perceptual distortion assessment

The mode decision is where RDO occurs, searching for an effective means to represent a block, be it a single block, a series of sub-block or no residual information should be encoded. In turn, those candidates where distortion is perceptually undetectable have an increased likelihood of been selected. This could result in the VCL undergoing perceptual bit-redistribution, where larger

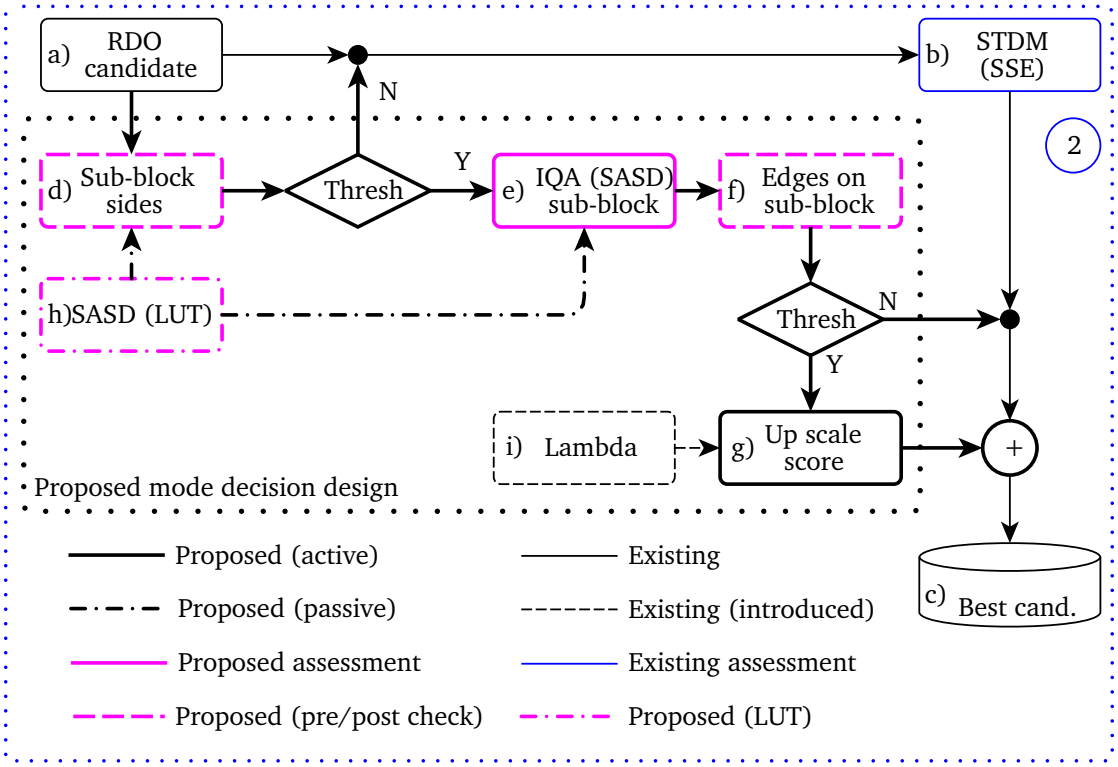


Figure 7.3 Proposed mode decision design

sub-blocks are applied for perceptually homogeneous regions and smaller blocks for perceptual significant areas, to retain the perceptual integrity of the video. The proposed mode decision reflects the common design overview, a pre-check, perceptual assessment and a post-check as illustrated in Figure 7.3. This allows for more observations to pass at the initial pre-check stage, with greater scrutiny applied on the post-check following the perceptual assessment. In this case, the post-check, is where the 1/4 block edge test is applied on the perceptual map of the respective sub-block. From the modelling this approach produced a high proportion of observations through the perceptual route, however, these were largely where 1-SSIM scores were high. Another feature which is not shown, yet part of up-scaling the final score, is the minimum SASD score threshold, this acts as a tolerance region before appending the perceptual cost. This threshold check is an important feature because in mode decision choices are few, thus when the perceptual difference is low, the additional cost is not applied. This means that any perceptual scores applied, lies between the minimum threshold for the given block size and a capped limit of 1/8 of the SSE score.

7.2.4 Proposed prediction perceptual distortion assessment

Since prediction is the initial stage it has greater variety and volume of candidates to evaluate, this means that any solution must be complexity friendly. For that reason, the proposed prediction IQA workflow has a double pre-check threshold prior to the distortion assessment, this is similar to the proposed rate-control design, and unlike rate-control in predication applies APC to keep the complexity low. This is shown in Figure 7.4, where LUT is used directly to gather APC scores. Compared to the proposed rate-control where ppwAPC is used, this is a lot simpler. However, the use of a double threshold prior to the distortion assessment dramatically reduces those candidates for the perceptual route, as shown with the modelling and simulation results. Consequently, this design limits the potential for additional complexity on the encoder.

7.3 Challenges for implementing LUTs

Underlying this low complexity implementation for PVC is the use of LUTs, as a means to minimise the additional complexity required to perform IQA.

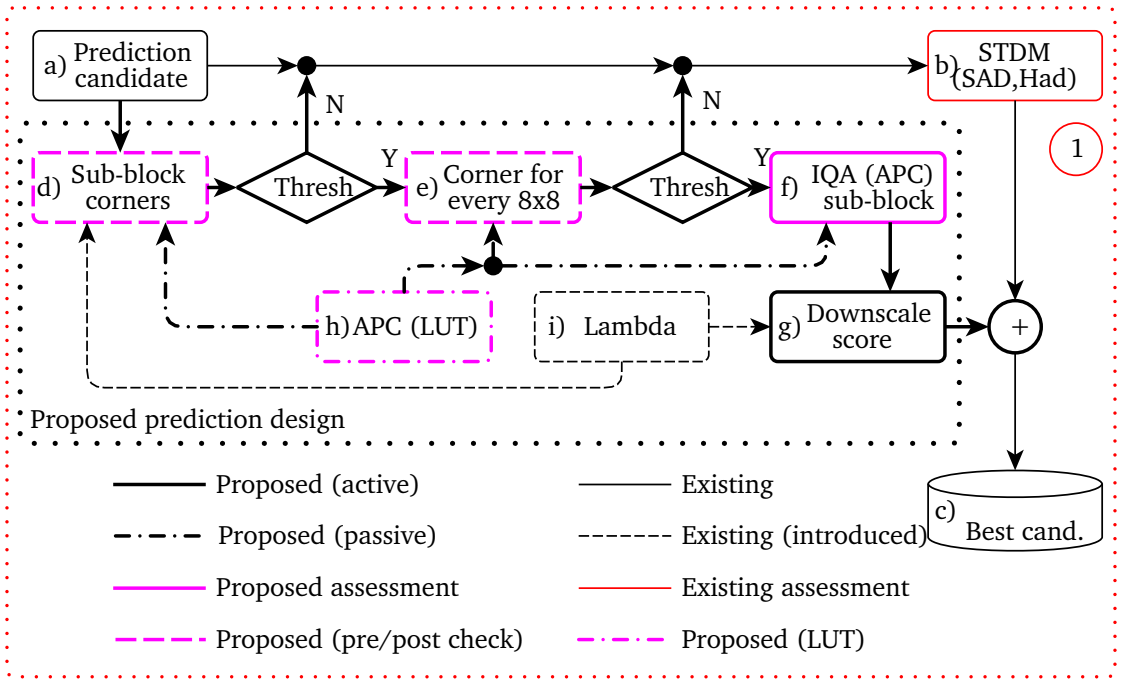


Figure 7.4 Proposed prediction design

Since the IQAs used are pixel-based, they are access via arrays and the scores are accumulated. However, these LUTs vary in size and means of access by the respective hybrid STDM-IQA framework workflows, as the LUTs are designed around the respective STDMs for which they operate with. Each of these will be described in this section.

7.3.1 Using look-up-tables (LUT) to save complexity

An LUT is an array of values in memory to save computational processing, it is used in this implementation to store the respective perceptual algorithm scores for the various combinations of original and reconstructed pixel values. While this suggests a 2D array, it is more efficient to access a 1D array. Since the pixel range for an 8 bit is 0 to 255, this means each row can be accessed by a fixed shift of 2^b, where b is bit depth equal of 8. Furthermore, because the pixel-based IQA can be calculated alongside existing STDM, then the LUT can be arranged as rows for original and columns for differences. The efficiency can be further improved with the LUT can be arranged to be zero-difference-aligned as shown in Figure 7.5, where the rows are the original value, while the columns contain

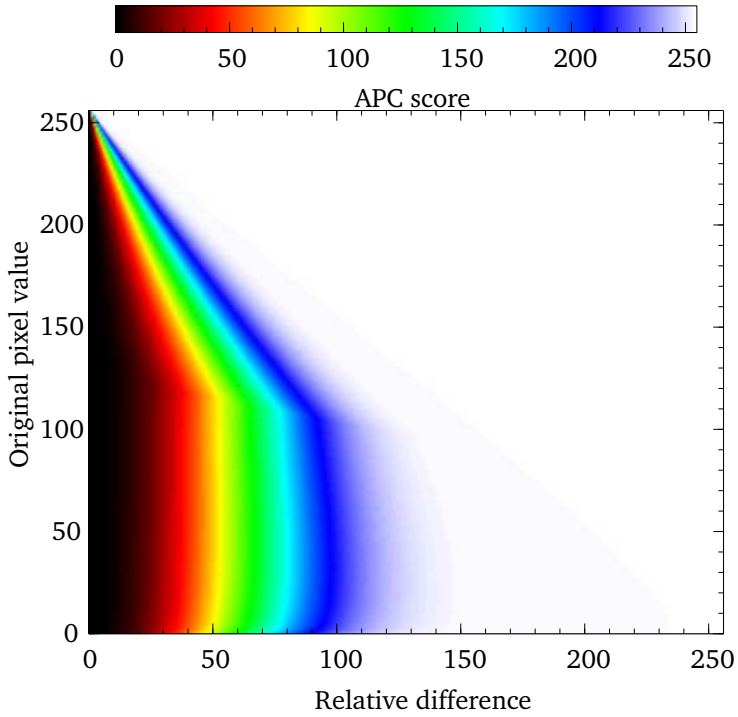


Figure 7.5 Zero difference aligned APC LUT

the relative differences to the original. In terms of operational speed, the zero-difference-aligned LUT is able to locate the values quicker where the differences are minor.

In SAD the differences are put through an absolute function and so are always positive, for SSE and Hadamard the differences are not processed and so can be negative. To avoid applying an absolute function, the LUT can be doubled in width by reflecting around the zero-difference column. This means that every entry must start half way along the row, however, as this is fixed in the implementation, the operational speeds are maintained. In all, three LUTs are required, one as shown in Figure 7.5 for SAD, a reflective version of Figure 7.5 used during rate-control (as well as sub-integer prediction), and a reflective zero aligned LUT based upon SASD for mode decision. These LUTs save on computational time, yet to store these LUTs during encoding requires additional memory usage. As these LUTs are designed to store 8 bit values, each entry can be stored as 8 bit integers. This means that each LUT will occupy memory (of data type size × array width × array height), for APC this is $2^8 \times 2^8 \times 2^8 = 2^{24} = 16$MB, while for the reflective LUTs it will be double this, at 32MB each. Therefore, the total memory required to store the 8 bit versions

of the proposed LUTs is 80MB. Given that potential complexity savings by using LUTs this additional memory cost is acceptable. However, extending using LUTs for higher bit depths of 10 and 12 bit, will increase the LUTs size by 4 and 16 times respectively, this means some form of optimisation is required to support higher bit depths.

7.3.2 Adapting rate-control to access LUT

The perceptual activity assessment is based upon the Hadamard 8x8 activity assessment on intra slice, which limits itself to the beginning of a GOP or intra frame. For the existing activity assessment the original 8x8 pixel array is assumed to be the differences. For perceptual activity, the process of gathering is a weighted form of APC called ppwAPC and done via single in-line function, this is because pixels can have several corresponding horizontal and vertical Hadamard transform pixel pairs. The use of ppwAPC is used as part of the pre-checks and later in the IQA, this means that when full IQA perceptual activity assessment is undertaken, only the internal 6x6 inner square of the sub-block is required. Therefore, despite ppwAPC being complex, the regulated use via the pre-checks, minimises the complexity load where possible.

7.3.3 Validating variables before accessing LUT for mode decision

The proposed IQA for mode decision requires prediction candidate information alongside the differences so that the LUT can be accessed. This means, encoding intra and inter block searching must pass the prediction pixel arrays during calls for perceptual assessment. Also, these prediction candidates must be tested to ensure they are valid and suitable, since they could duplicate the reconstructed or be a null pointer. For this reason, the pixel array should be boundary tested to establish whether it is needed, in this case the initial and last pixel of the prediction against the original pixel array.

7.3.4 Using two different LUTs in prediction

The proposed perceptual assessment during prediction is largely similar to rate-control, however, this relies on a lot fewer pixel points. This does introduce risk of incomplete coverage as shown during modelling where 5% of total observation had high 1-SSIM scores. Prediction has the highest volume of observations more so due

to AMP, this means that low complexity is an important design factor. As the block sizes can be non-square and varied in size, a pixel base solution can be adapted with the similar levels of overhead. Unlike the other two stages which manage the additional complexity by involving non-perceptual calculations during perceptual pre-checks, prediction does not do this. This means that perceptual pixel values during pre-checks are not used in the perceptual score. Fortunately, the overhead for the double pre-check is low, also two different LUTs are used meaning that likelihood of performance degradation due to race condition is lowered. For pre-checks the reflective LUT is used as an absolute function has not being used, while for the assessment SAD has been calculated and so non-reflective LUT is called upon.

7.4 Design of proposed hybrid subjective testing

Implementing a new proposed sub-block level PVC solution requires evaluating the effect of its perceptual redistribution, particularly on the intended audience of the HVS. This involves conducting subjective testing to explore the preference of the proposed PVC solution or the original. Existing subjective testing is able to provide a means to evaluate whether one encoder is subjectively preferred over another, or provide a measure of an encoder against an uncompressed video However, these approaches do not offer a relative measure to each other that is randomly ordered. Being able to measure the relative preference directly means that the same participant is able grade a pair of encoding sequences relative to each other than to a reference. A relative grading can provide magnitude and a direction of preference, which is a higher resolution of understanding than what existing methods of subjective testing provide. Also, randomising the order can avoid bias for the first or second video sequence being shown. In this section a proposed subjective testing method will be described to provide a relative grading based upon existing subjective testing methods.

7.4.1 Proposed DCR-PC subjective testing

Subjective testing was undertaken as a means to evaluate whether perceptual redistribution have a lower subjective score than existing method. This is because previously, when prediction only was modified using SSIM in H.264/AVC the

picture quality was observed to be much lower under the configuration of low delay P, however, this was not subjectively tested. Since this is a PVC solution at each front-end stage with its respective IQA workflows, subjective testing will provide an overall measure.

Video testing standards support different subjective testing methodology, among which include pair comparison (PC) and degradation category rating (DCR) (ITU-T,

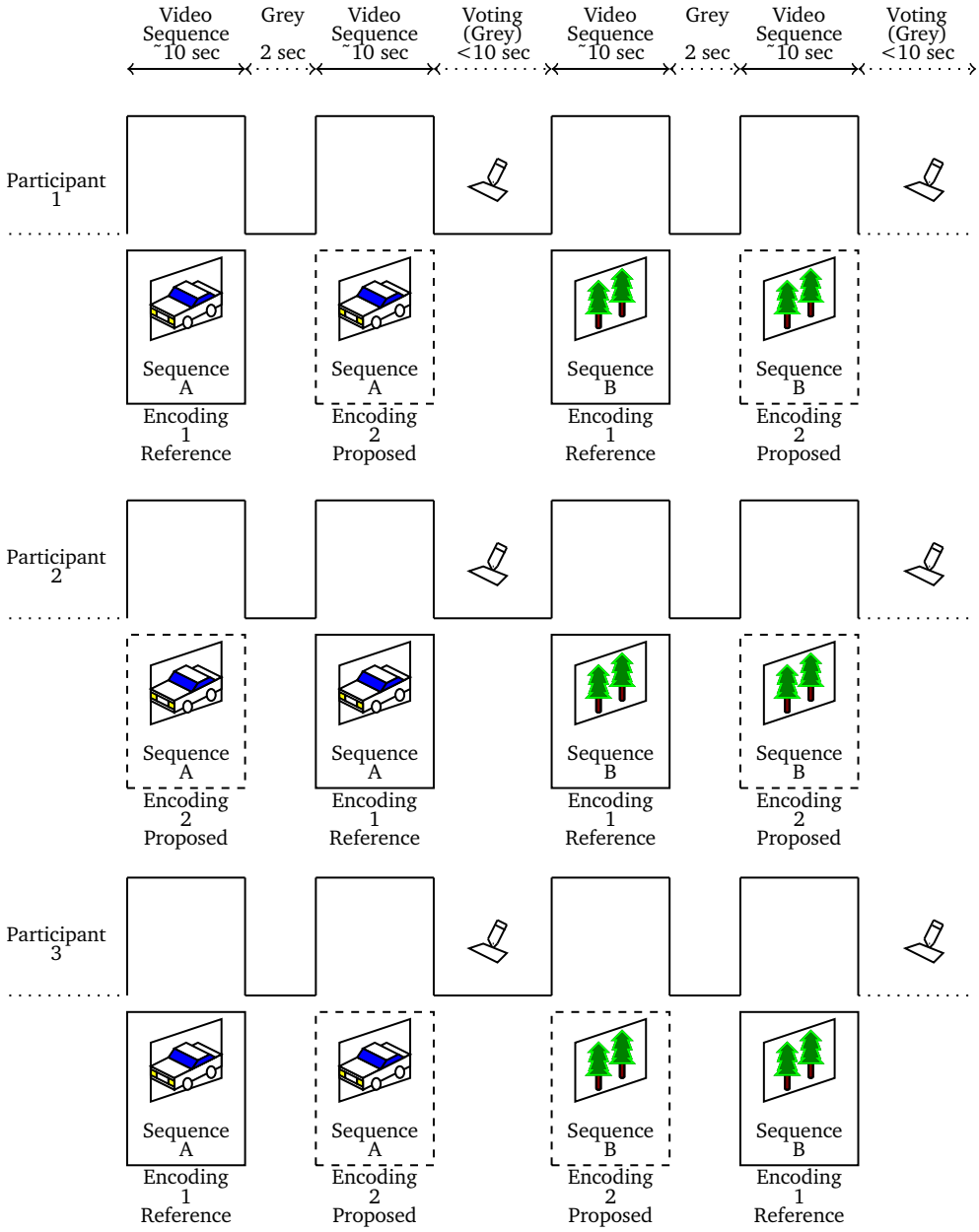


Figure 7.6 Proposed hybrid pair comparison (PC) degradation category rating (DCR) method

2008). PC encourages randomising which respective encoder is shown first, while DCR suggests a five-level rating scale. A hybrid of these existing subjective video testing standards is proposed, combining PC and DCR, which produce a subjective testing methodology that is less prone to bias and potentially more descriptive of any potential differences. This proposed hybrid method is shown in Figure 7.6, where each participant is given a different pair orders of encoded video sequences to assess. Consequently, this does mean that greater preparation is required in creating respective order of video sequences and additional time is required after completion to normalise results. In terms of data collating and transforming, an on-line form was used which saved the results to a spreadsheet, then a transform was applied to retrieve the ordered results. To keep the on-line form user-friendly, only four voting questions were presented at any one time, each with a multiple choice answer available. An example is shown below:

(Video sequence title)

Please select whether if any subjective difference was experienced

- Mostly preferred 1st video
- Slighty preferred 1st video
- No difference
- Slighty preferred 2nd video
- Mostly preferred 2nd video

Each of the video sequences in Table 7.3 were shown across a range of bit rates except for the 1Mbps encoded video sequences. It has been acknowledged that participants undertaking subjective testing are liable to suffer fatigue and lose concentration beyond 20 minutes of subjective testing (Pinson, Janowski and Papir, 2015). Ethical consent was not formally requested, as risk to the participant was considered to be minimal, however, verbal consent was asked and a handout was provided when recruiting participants, explaining the experiment as outlined in (ITU-T, 2008). To make the participant familiar with the experiment a separate example of a two video sequence test (using CrowdRun and BasketballDrive video sequences) and an example on-line form was shown before the full test. Ishihara

test for colour blindness was not conducted nor were participants asked if they wore corrected vision for viewing monitors as all were active computer users and enrolled from their desks where they worked on a computer.

7.5 Hypothesis for testing methodology

As explained earlier in the introduction, the implementation is a means to validate the design, however, it is through testing which verifies whether the concept is operating as intended. This means that the testing methodology should reflect an understanding of the proposed hybrid STDM-IQA framework and evaluate its limitations. From the initial set of experiments based upon hybrid SSIM-SATD PVC solution, only objective testing was applied. Subjective testing on that experiment would have allowed a base line to be created, however, the objective results suggested it was not worth pursuing. In the previous chapter, a new tool was created called visual VCL tool in order to visualise the effects on the bitsteam and to simulate of the proposed IQAs within the proposed framework. This means the visual VCL tool can be used to evaluate, the proposed encoder by way of encoded bitstreams and analyse whether perceptual bit-redistribution had occurred.

7.5.1 Hypothesis for objective testing and visualising VCL

The VCL illustrates the partitioning within the frame as a set by RDO process of block matching during on intra and inter coding. For the existing encoder, all changes are treated equally, however, the HVS is sensitive to brighter lighting regions and where textures or boundaries are present. The proposed encoder is designed to produce higher scores for candidates which are perceptually significant, thus encouraging the RDO to apply smaller sub-blocks in those regions. As the previous attempt of SSIM based sub-block prediction resulted in higher distortion and poor video quality, this experiment is whether distortion can be increased with low or no perceptual loss in video quality. This will indicate that distortion in video can occur without affecting user experience given low complexity techniques. The proposed pixel-based IQA is designed to minimise the losses experienced under low delay P with SSIM based sub-block PVC. The reason for this is that the pixel-based IQA is less likely to experience the issue of statistically same candidate that affected SSIM based sub-block PVC. This means that where pixel IQA is triggered it should

offer more computationally competitive and perceptually accurate response than the previous SSIM based sub-block PVC. Equally, with the use of the VCL tool, the changes in perceptual redistribution should be made visible.

7.5.2 Hypothesis for subjective test

In the previous findings where SSIM based prediction was implemented, the objective results were substantially poor under low delay P and this affected the overall video quality. As such under this experiment, the hypothesis was that the proposed is inferior, than the existing encoder. While the null hypothesis was set as that no differences between existing and proposed encoders would be detected. This means that a one-tail t-test on pairs was used to analyse the results. A one tail test is considered a stronger test than a two tail test as the probability is concentrated to one side than being split between two ends of the normalised bell curve. The threshold in this one-tail was set at 5%, this means that if normalised probability fell into that 5% then it would deem the proposed encoder to be subjectively inferior to the original HEVC encoder.

7.6 Testing methodology for experiments

The proposed hybrid STDM-IQA framework consists of modifying the rate-control, mode decision and prediction, based upon a common design concept, however, the technical challenges faced extend to supporting the IQA path. The proposed visual VCL tool and hybrid DCR-PC subjective test have been produced to enable testing of the proposed PVC solution at the sub-block level and by the HVS respectively. The implementation involved seeking a low complexity solution and coverage of all sub-block sizes, especially as HEVC supports AMP. While modifications to the decoder enabled the visual VCL tool to be created, it also allowed the calculation of SSIM per frame and block type/size cost in bits per frame. This meant that frame level perceptual quality and bit-redistribution can be measured, which is critical in understanding whether sub-block level PVC affects the VCL and what impact it has perceptually.

7.6.1 Implementation

The implementation was undertaken on HM HEVC version 16.6 reference encoder (JCT-VC, 2016). This codebase was modified whereby the respective

hybrid STDM-IQA design was used to create the low complexity in-loop PVC solution. The proposed encoder was developed using an integrated development environment called Qt Creator (Project, 2015). Since the LUTs were defined as zero-aligned, this enabled the existing STDM calculation of the differences to be applied as the offset along each array. Due to the high number of supported sub-block sizes in prediction, including AMP support, this dramatically magnified the source file. The LUTs were limited to 8 bit as 10 and 12 bit LUTs would be excessively large in source and runtime memory. In additional, at the time, video sequences were limited and these were full HD with low dynamic range of 8 bit. Also, development was limited to three video sequences, RaceHorses, CrowdRun and BasketballDrive, the last two being HD. To keep independence between development and testing, these video sequences were excluded from those made available to evaluate the proposed encoder.

7.6.2 Testing overview

The proposed PVC underwent testing on video sequences representing different applications. The number of video sequences available for testing was limited and encoding the full video sequence can take substantial amount of time. For that reason, the video sequences were encoded in one of two configurations, random access or low delay P as shown in Table 7.3. Random access was chosen for VoD or off-line/pre-recoded scenarios. Low delay P was selected to reflect on-line live streaming or two-way communications where low latency is required for reasons of responsiveness and/or limited processing capabilities exist. These video sequences were classed as either communications or monitoring. The communications label is broad, as it can be both one-way and two-way depending upon the context, while monitoring is designed for remote surveillance, without engaging with an audience. Since no single video was in both configurations it meant that any common observed traits could be associated with the configuration, with the encoder or general video content. Furthermore, the video sequences were encoded at five set bit rates of 1, 2, 4, 8 and 16 Mbps, which representing limited bandwidth, through to broadcast and off-line media.

	Communications		Monitoring	
Random access	ParkScene	Tennis	PedestrianArea	Riverbed
Low delay P	Kimono	DanceKiss	FlagShoot	BQTerrace

Table 7.3 Videos sequence testing matrix.

7.7 Experiment set-up for objective testing

The video sequences were encoded on a system with an Intel iCore 7 920 processor with 7 GB of RAM running Ubuntu 15.04. To avoid issues of heat affecting performance, a single video encoding was run at any one time. Also, to minimise other processes running the terminal emulator (tty) was used at the graphical login. Then a batch script called the respective encoder at each of the five bit rates for a given video sequences. Finally, at the end of test run the computer was switched off and allowed to cool-down. This protocol was repeated per video sequence for both original and proposed encoders. Both the encoder and decoder analyser applications produced logs, these were analysed for performance (time), image quality (PSNR and SSIM), bit distribution (bit-usage by block size/type). These results were considered as whole on whether if for similar picture quality would remain if bit-redistribution would occur.

7.7.1 Visual inspection of the VCL

As part of the batch script, the modified decoder was called to extract a specific frame in the video sequence, in this case frame 77 from the video bitstream. The choice to extract frame 77 is related to random access configuration, which has a long GOP of 32 frames each. As such frame 77 allows the encoder to settle with two GOPs, then the hierarchical sub-GOP structure is applied. The hierarchical sub-GOP structure reoccurs every 8th frame and is shown to be a means to balance picture quality and bit usage (Hong et al., 2010). At frame 77, the frame relies upon five other frames, making it a highly compressed frame with one of the lowest bit usage within the sub-GOP and with a higher likelihood of distortion being present. Therefore, the choice of frame 77 as a means to conduct visual assessment is related to demonstrate VCL changes visually.

7.8 Experiment set-up and design for subjective testing

To implement the hybrid PC-DCR test sequences the respective HEVC bit-streams were collated under the Matroska (MKV) video format wrapper (Matroska.org, 2015). These videos alongside videos representing periods of grey (used for intervals and voting) were merged using a tool called mkvmerge (Bunkus.org, 2015). The playback was under a media player called video lan client (VLC) (Videolan.org, 2015). The subjective experiment was conducted with a Sony 40" TV (Model KDL-40EX1) at 1080p resolution which was calibrated using AVS HD Rec. 709 (AVSForum, 2016). Finally, the video sequences were shown on a Dell Latitude M4500 laptop running Ubuntu 15.04 on an iCore 5 with 4 GB memory. Ubuntu 15.04 was used because under Windows 7 32 bit, the playback was poor and under Windows 10 64 bit it was not consistent. However, even under Ubuntu the desktop manager was substituted with MATE to provide a 20% headroom on processor usage. Also, to ensure smooth playback the video sequences were played from an external solid state drive (SSD) via a USB connection, thus avoiding any potential lag due to the internal mechanical hard disk drive (HDD). The subjective test was run inside a small-medium sized meeting room, where participants would be positioned from the TV such that the distance is three times the TV height. This set-up was in-line with recommendations of BT.500 for best practises for subjective testing (ITU-int, 2012). The experimental overview design is shown in Figure 7.7, where it states that the encoded video sequences were tested on a TV. This diagram is expanded upon in Figure 7.9, which is inspired by (Kilkenny et al., 2010).

7.8.1 Lack of experiment on mobile phone/tablet

Ideally, the proposed solution is designed to operate within portable or low powered devices. However, the subjective testing could not be conducted on a mobile phone due to the lack of smooth playback of the encoded HEVC bitstream. Similarly, running uncompressed video would require more storage space than available and then demand high throughput rates between storage and video rendering. This means that unlike laptop processors, where HEVC decoding can

run in software, to conduct this experiment under a mobile phone or tablet, HEVC decoding requires hardware acceleration.

7.8.2 Experimental design

The experimental design shown in Figure 7.9 is designed as a representation of the process by which subjective testing was undertaken. The experiment itself required that participants be randomly allocated a number that corresponded to the randomised encoder order per video sequence. A total of eight video sequences were used, and a uniform, probability of success was selected, meaning p = 0.5. The binomial probability distribution figures were then scaled up (by 30) and rounded as shown in Figure 7.8. This Gaussian bell curve distribution as represented by the line graph highlight accumulates to 30, stating that 30 variations of where the encoder order are required. While, the stacked bar graphs reflect the required proportion of randomisation per given trial. This means that the order of these eight video sequences remain constant for each trial in the subjective test, however, the order of which encoded video sequence pair is shown (A then B or B then A) may differ within and between trials. For example, where the trials is three, the distribution states that there should be seven collections of the eight video sequences In each of these seven collections, three of the video sequences should show the encoding by encoder A then that of encoder B. Similarly, with trials at six, three collections are required in which encoder A is shown first for six out of the eight video sequences. Overall, this meant that these subjective testing trials are less liable to bias due to this level of preparation.

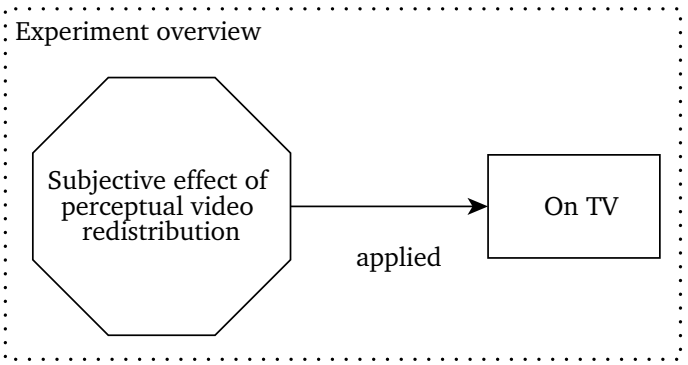


Figure 7.7 Experiment overview

In the experimental design in Figure 7.9, the two variables with which the video sequences were been evaluated for were configuration and bit-rate. The term configuration referred to random access and low delay P, which reflect types of applications described in the critique. While the use of bit-rate was to be representative of the variety of bandwidths choices that need to be supported, ranging from 2Mbps to 16Mbps. The analysis was based upon relative impairment and since the encoding sequences were shown following each other, analysis was performed using paired t-tests. Finally, the experimental design shows that the analysis considered the two variables of configuration and bit-rate, examining for whether subjective test recognised any changes. The previous SSIM at sub-block experiment demonstrated a loss in objective measurements. Here the analysis was set to examine whether this loss was replicated in subjective with the proposed hybrid STDM-IQA framework.

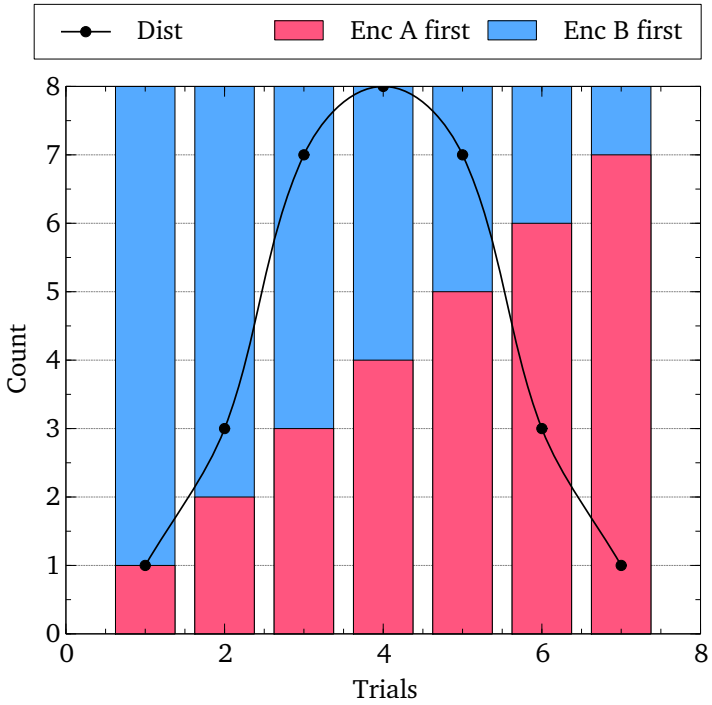


Figure 7.8 Subjective testing distribution of trials and for each trial the number of occurrences where encoding A or encoding B is shown first per video sequence

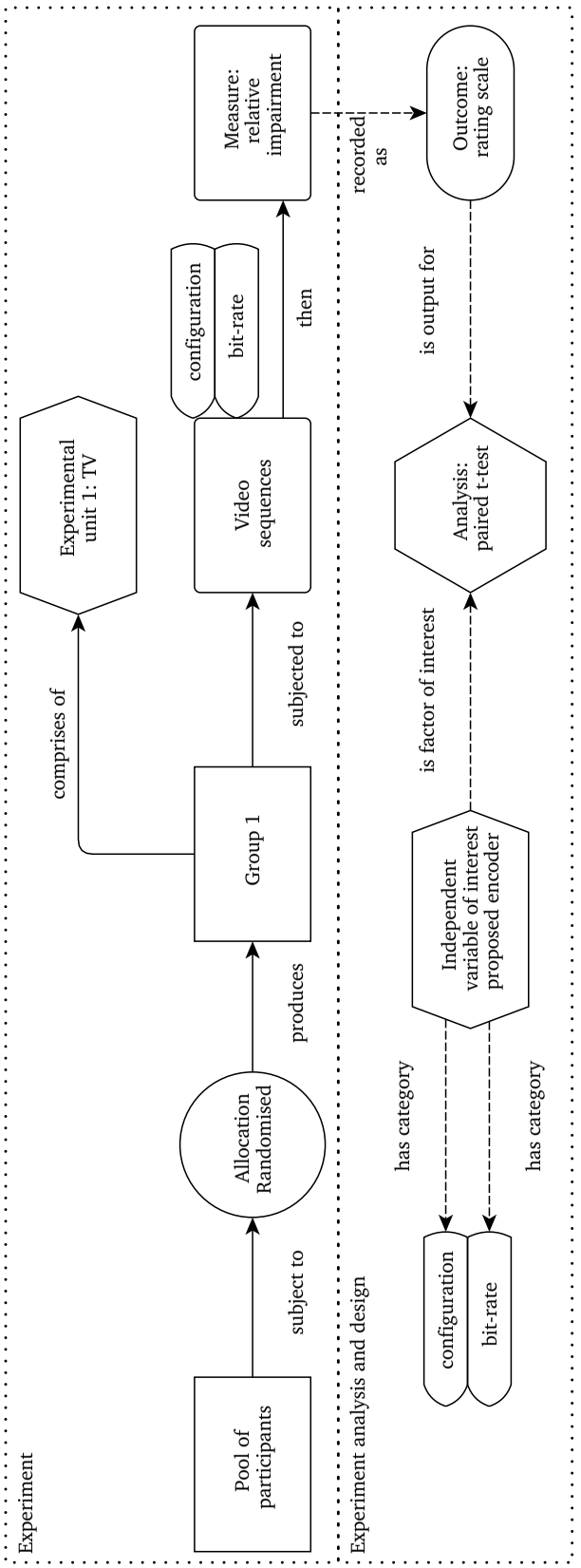


Figure 7.9 Experiment design

7.8.3 Subjective testing experiment

The experimental design highlights how the results will be analysed by encoding configuration (random access and low delay P) and by bit-rate (16, 8, 4 and 2 Mbps). As the experiment consisted of paired results, they were analysed as t-paired tests for their relative impairment. Overall, the participants were shown each of the video sequences at the highest possible bit rate (16Mbps), then three subsequent runs at half the previous bit rate (8, 4 and 2). The experiment involved 11 participants, however, the videos were repeated with only one variable change, in this case of bit-rate. This means that even with limited number of participants, the repeated measures would minimise or exacerbate variation.

Furthermore, the measures of evaluation were broader than individual videos, and set by configuration of encoding profile or bit-rate. This allowed diversity of the video content to affect the experiment outcome, which means any change should be significant as it would be averaged across a collection of video sequences. The use of grouping results by their respective bit rate and configuration, meant that 44 observations were gathered per grouping. This is because each configuration had four video sequences, four for random access and four for low delay P, times 11 participants, for each of the eight groups. These number of observations, 44, satisfy the central limit theorem, which state that beyond critical mass of around 30, additional observations tend to stabilise, reducing the benefit of capturing further observations.

It is acknowledged that with greater observations, a higher resolution of analysis is possible to the individual video sequence, however, this was not the scope of this experiment. The experiment was to evaluate whether a trend by bit-rate or by encoding profile existed to suggest that the proposed encoder was relatively inferior. On completion of the 20 minute of video sequences, the on-line form would ask participants for their gender and which age group they were. Of the 11 participants six were male and five were female. While those who shared which age group they were; six chose between the ages of 25 and 34, three stated they were between the ages of 35 and 44, finally one indicated they were between 55 and 64.

7.9 Chapter summary

Following on from the previous chapter where the proposed hybrid STDM-IQA framework for the respective pixel-based IQAs were presented, this chapter implemented this design into HEVC, a hybrid block-base encoder. HEVC presented its own set of technical challenges to allow PVC to occur. Following a common proposed design pattern of pre-check(s), IQA and post-check, the designs were developed further for the respective front-end stages accordingly. Subsequently, the remaining chapter described the proposed testing methodology that was applied. In particular, the choice of video sequences, their encoding settings were presented. Then measurements used to evaluate performance including using the proposed visual VCL tool and a strategy for conducting subjective testing was presented. This sets up the premise of the next chapter, which will contain results of the PVC proposed implementation against the reference encoder.

Chapter 8

Results for low complexity in-loop PVC in HEVC

In this chapter, the proposed low complexity in-loop PVC HEVC encoder will be evaluated relative to the reference encoder. This will be measured using objective, visual and subjective methods. Together, they will aim to measure whether PVC during sub-block candidate selection leads to bit-redistribution that retains perceptual integrity and is complexity competitive. In order to be representative of today's range of video environments and content, five bit-rates were chosen to run on two different encoding profiles, each encoded with four video sequences. This means the test material and set-up is a fair representation of applications described in 'Applications for low complexity PVC' in Section 3.1.

8.1 Objective testing results for video sequences

The objective results are shown in two parts, from the encoding logs, and by modified decoder logs. The encoder logs provide PSNR and encoding time, while the modified decoder calculated the averaged SSIM and SSE scores per frame. Across both configurations, in Figure 8.1a and Figure 8.1b, the encoding

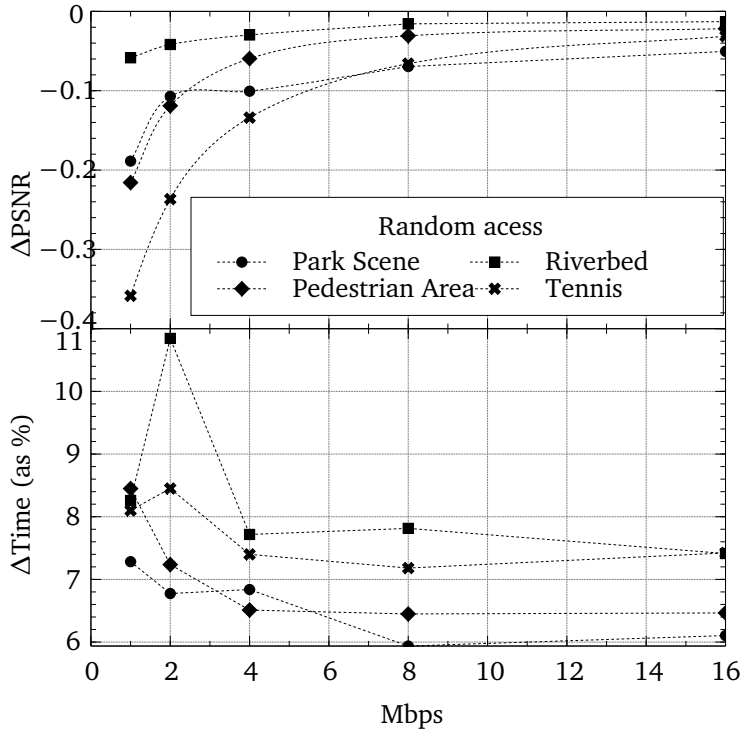
timing results fluctuate between +6% and +11% for the proposed encoder. This is reflective of the increased activity of for the proposed hybrid STDM-IQA framework undertaking the IQA path. Generally, the timing increase was higher for lower bit-rates, especially at 2Mbps. This increase in timing at lower bit rates could be related to the higher likelihood of distortion with abrupt changes which would exceed the pre-check(s) threshold and cause pixel-based IQA to be calculated. Looking at these results by configuration, Figure 8.1a and Figure 8.2a respectively show that for random access the ΔPSNR losses reduced as bandwidth increased, while for SSE and SSIM the R-D curves were not too dissimilar. For the equivalent graphs under low delay P in Figure 8.1b and Figure 8.2b illustrated virtually no changes. This suggest that the IQA path is being applied in random access and that little or no change is occurring in low delay P.

8.2 Bit usage by sub-block type/size

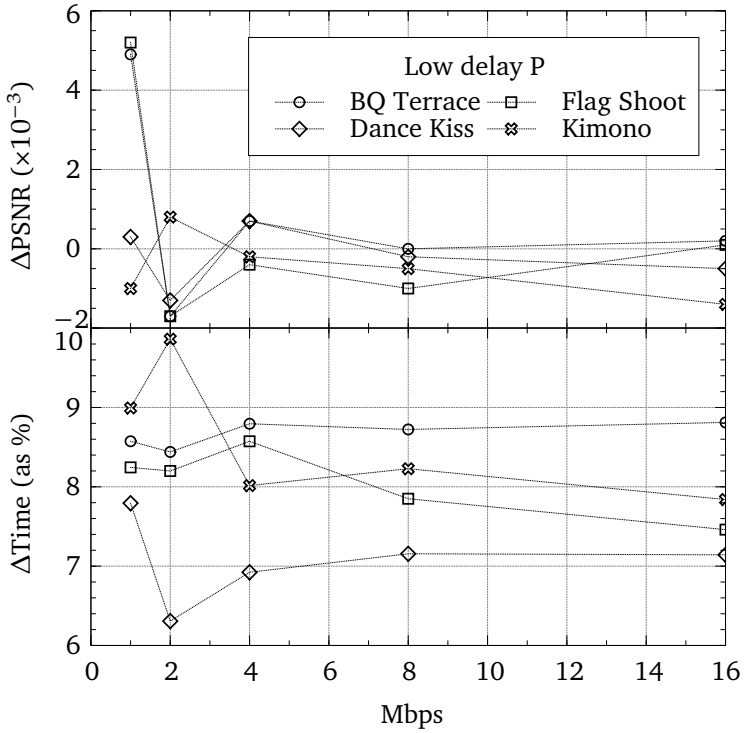
The initial objective results showed limited overall disparity between PSNR and SSIM, however, this is not evidence that the proposed PVC solution was affecting the choice of blocks types/sizes. The results shown here use the logs from the decoder analyser of the modified decoder analyser to gather the bit usage by block type/size by sequence for a specific frame respectively. By being able to show bits are distributed by sub-block sizes and type, it should be possible to describe the decisions being made by the proposed encoder.

8.2.1 Video sequence percentage averaged across bit-rates

Using the decoder analyser logs, it is possible to gather bit distribution per decoding. As PVC adapts to the nature of the video content, these results are presented by video sequences. This means the percentage bit usage differences can be averaged across bit-rates. In Figure 8.3 the relative percentage bit usage differences are shown as bar charts per block size/type. In the graph, low delay P encoded sequences are shown above and random access below in Figure 8.3 for the same x-axis of block type and size. In terms of the y-axis which represents Δbit-usage, for low delay P compared to random access is equivalent to $1/1000^{th}$. This means that low delay P exhibits little or no perceptual bit distribution across the whole video sequence. This is reinforced when a trend line, coloured in

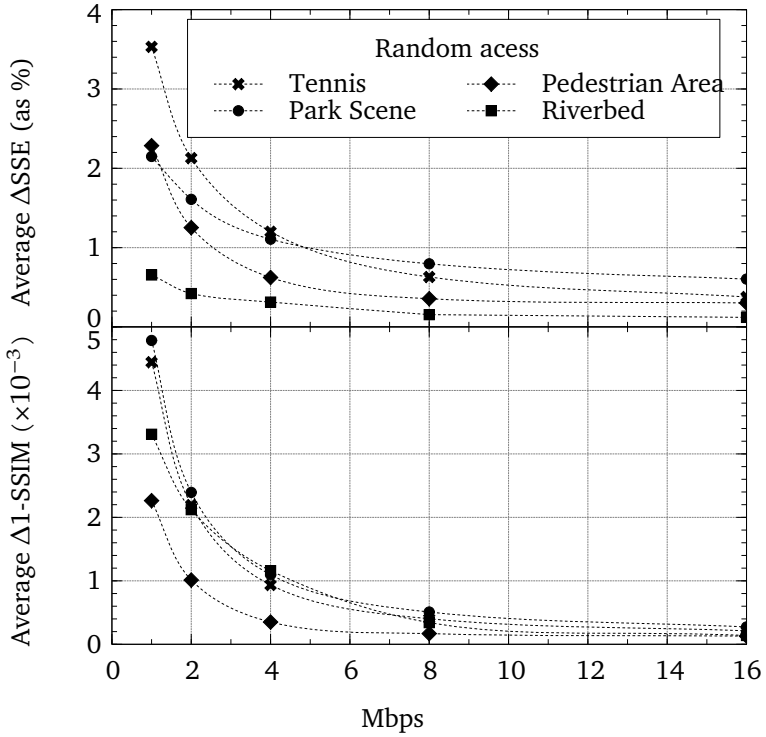


(a) Random access

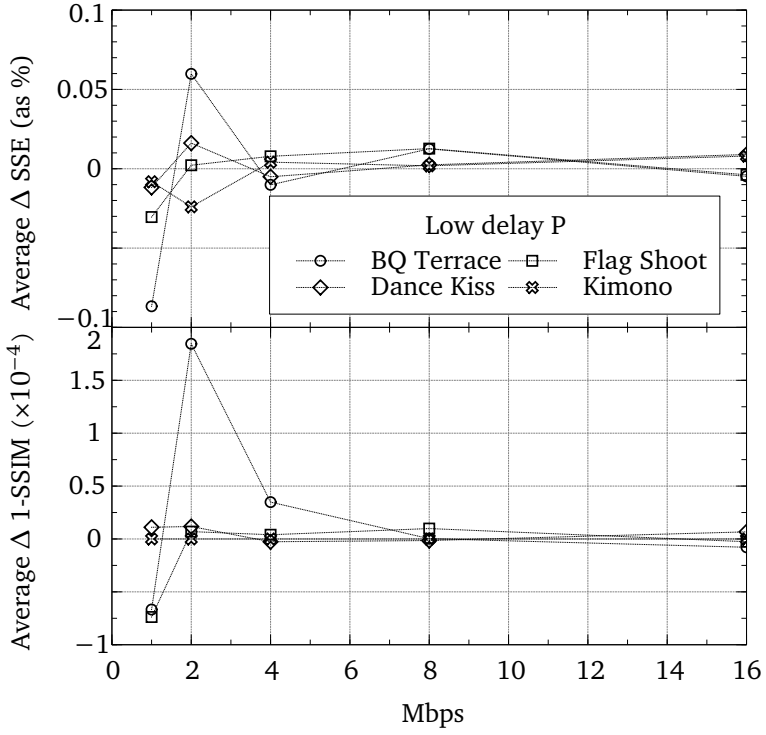


(b) Low delay P

Figure 8.1 Results: Rate Δ distortion curves non-perceptual and perceptual by configuration



(a) Random access



(b) Low delay P

Figure 8.2 Results: Average Δ SSE and 1-SSIM graphs by configuration

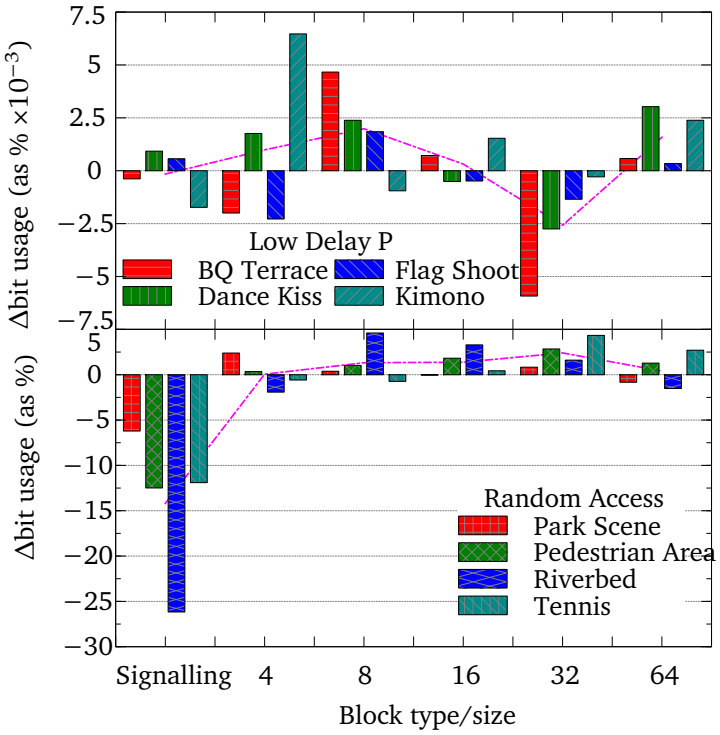


Figure 8.3 Decoded video Δbit distribution by block type/size for low delay P and random access configuration, averaged across by per video (bar charts) and by configuration (line).

magenta is used to represent the average for each configuration. Under low delay P, once the trend line is considered in relation to the scale, virtually zero bit-redistribution occurs with the proposed PVC solution. While for random access, the trend line in relation to the scale shows significant bit usage changes, with signalling information being reduced and increases in medium sized block widths of 8 to 32. This occurs mostly on Riverbed which is a highly active video, then PedestrianArea and Tennis which are moderately active.

8.2.2 Individual frame for 1 and 16 Mbps

The above results suggest that the proposed PVC solution significantly changes the bitstream of video sequences encoded under random access, both PSNR and bit usage distribution are affected while SSIM is maintained. In Table 8.1 the bit distribution of an individual frame is examined at both low and high bit-rates of 1 and 16 Mbps respectively. These extreme bit-rates were selected to understand whether the average bit usage differences per video sequence is a fair reflection of the actual bit usage differences by frame. The results were gathered by modifying

the decoder analyser to go further and obtain bits distribution per frame. The same frame across all video sequences was chosen to ensure consistency. In this case, the frame 77 was selected as it was beyond the first GOP and allowed the encoder to settle down. The results of frame level relative to bit usage by block type/sub-block size are presented in Table 8.1a for random access configuration and Table 8.1b for low delay P. These results show substantial variations in the averaged full sequence encodings are not reflected in the individual frame level. This suggests that other bit-rates or adjacent frames may have substantial differences that can reflect these average values seen in full sequence bit usage results. In addition, the results that random access and low delay P both exhibit bit re-distribution of similar scale, with random access encoded streams having a greater magnitude. The level of perceptual bit re-distribution is higher at low bit-rates, however, in all low delay P is likely to have lower levels of perceptual bit re-distribution. Therefore, the results show that activity is occurring for both configurations, however, for low delay P it is less likely which results in virtually zero bit re-distribution change at the sequence level.

8.3 Overview of visual VCL tool for encoded video sequences

The respective objective and bit usage figures describe that bit-redistribution is occurring, mostly with random access configuration at low bit-rates. However, it is important to visualise this redistribution, to understand whether the changes to the VCL are in-line with those related preserving the perceptual integrity of the frame.

The visual VCL tool shows partition and meta information (QP and bit useage) based upon the encoded bitstream and presents this as a heat map. Visualising the QP applied to residual is the most precise measure as each sub-block can be individually highlighted. For bit usage, each LCU is highlighted based upon the accumulation of sub-blocks within each block, this is less precise than QP, however, this includes bits allocated to signalling, such as motion vectors.

Another feature which is offered with the visual VCL tool is the ability to simulate assessments and the proposed IQAs as part of the framework. These frames should

be considered by methods of assessment using an 8x8 pixel array, as these can assist in gauging how to interpret the proposed encoder. The frames will undergo simulation under rate-control for activity to understand bit budget allocation. Also, the frame will be perceptually assessed based on SSIM to understand how the proposed behaved compared to the original.

In these heat maps, the colour scheme used is where red is high and blue is low, like the first five colours of the rainbow, red, orange, yellow, green, blue. These heat maps also have a series of red or white lined boxes which have been added at the same positions per video sequence to illustrate and compare specific regions between different visual VCL results. Where these heat map use grey scale, this means that no metadata was available suggesting in terms of QP that no residual bits, for LCU no zero bits stored, and for simulation no IQA related score applied.

8.4 Visual VCL frame QP distribution

The frame QP distribution for the respective video sequences are shown in Figures 8.4 to 8.11. The results show 16Mbps and 1Mbps of each video sequence with respect to the original and proposed encoder. The partition distribution and residue quantisation is shown by the overlaid grid and heat-map respectively. The use of square grid indicates the sub-blocks, 4x4, 8x8, 16x16, 32x32 and 64x64.

8.4.1 Random access

The first four figures, Figures 8.4 to 8.7, are where random access configuration was applied. Depending upon the nature of the content these can have some notable changes in their partitioning choices and where quantised residual should be stored. For where content changes are limited and has high activity like for ParkScene in Figure 8.4, the moving object, the cyclist can have more localised partitioning. In turn this allows, more static and/or homogeneous regions, such as the tree or park path, to be represented with larger sub-block sizes. The content for Tennis in Figure 8.5 is where the lighting changes are occurring due to the camera panning across the grill fence and by shadows on the tennis court. Where the definition is clear between foreground and background such as the individuals nearest the camera or the strong shadows, the partitioning tries to wrap around those objects. Conversely, when objects are indistinguishable between middle-

ground and background, the proposed encoder is likely to encourage larger block sizes, as shown by the couple near the middle of the frame.

The first two video sequences showed video, where under clear sunny day, giving large dynamic range and videos had isolated activity. In the next two videos of PedestrianArea and Riverbed, the lighting was either broken or under cloudy day, meaning limited dynamic range and videos had widespread activity. For PedestrianArea in Figure 8.6 shows the encoder struggles to determine the important regions. In terms of movement, a pedestrian's leg is more precisely tracked in the proposed than original as shown in the VCL. The middle-ground on frame consists of the pavement, which is largely static, is largely homogeneous, meaning larger blocks are encouraged. For Riverbed in Figure 8.7, the video sequence is a challenge for the respective encoders as local changes are occurring continually and simultaneously. However, the proposed encoder does attempt to use smaller block sizes where the sun light is being reflected off ripples of the river.

8.4.2 Low delay P

The following video frames, Figures 8.8 to 8.11, are those encoded with low delay P, from the decoder logs these were shown to have fewer changes in bit distribution. Kimono the previous results demonstrated it had the most differences in terms of bit distribution among the low delay P configuration. In Figure 8.8 for Kimono video sequence, in the foreground, the individual is tracked as they move across highly textured background, with little or no middle-ground. The partitioning and quantised residual tend to occur where boundary changes happen, such as the fold and overlap of the kimono dress or where pockets of light meet the branches. This allows for proposed encoder's VCL to offer more continuous sub-blocks of signalling or residue, which matches the content. For DanceKiss in Figure 8.9, the video sequence is taken in-doors, where the background is homogeneous and the movement is broad yet slow. The colours shown on the heat maps are all green indicating that the level of quantisation is very low in comparison to the other video sequences. This allows the encoder to preserve greater proportion of detail, as shown with the feathered scarf worn by the male actor. For the proposed, as seen with other VCL results, the partition structure seeks

to wrap around objects, in this case on the silhouette produced by the actor passing across the camera.

The next two video sequences, FlagShoot and BQTerrace are taken outside during the daytime. In FlagShoot, Figure 8.10, the proposed encoder demonstrates how bright homogeneous regions can be represented with larger sub-blocks compared to the existing encoder. For BQTerrace in Figure 8.11, the differences between proposed and original is restricted by the high amount of detail within the video content. Where these differences do occur, they tend to cancel each other out. Again, low delay P is unable to produce substantial perceptual redistribution, as the VCL residual QP illustrated, with minor differences in partition structure.

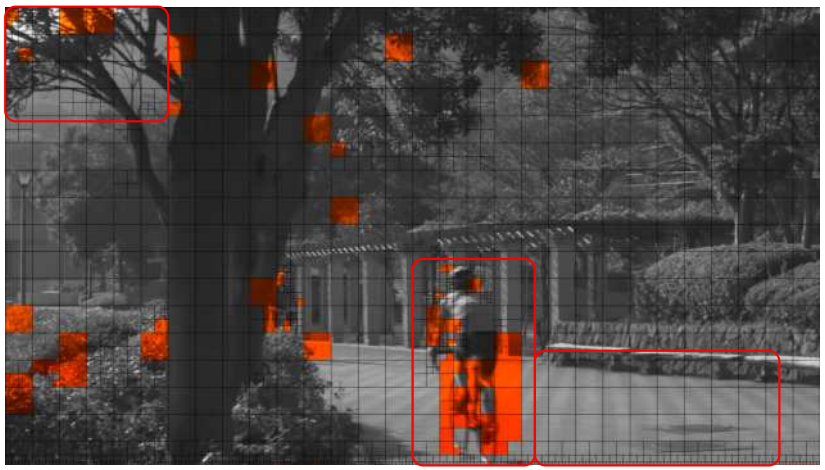
		Random access				
Video (1 Mbps)	Sig (%)	4 (%)	8 (%)	16 (%)	32 (%)	64 (%)
ParkScene	1.06%	0.18%	2.64%	-1.52%	-2.41%	0.05%
Tennis	4.74%	0.08%	0.49%	3.17%	-8.12%	-0.35%
PedestrianArea	5.52%	0.00%	-0.70%	-5.21%	-0.08%	0.48%
Riverbed	6.71%	-0.01%	-0.03%	-2.41%	-4.30%	0.04%
Average	4.51%	0.06%	0.60%	-1.49%	-3.73%	0.05%
Video (16 Mbps)	Sig (%)	4 (%)	8 (%)	16 (%)	32 (%)	64 (%)
ParkScene	-0.11%	-0.39%	-0.31%	0.61%	0.23%	-0.03%
Tennis	2.04%	0.24%	-0.57%	1.84%	-3.53%	-0.02%
PedestrianArea	2.49%	-0.60%	-0.41%	-4.82%	3.28%	0.06%
Riverbed	0.74%	-0.09%	0.13%	1.16%	-1.92%	-0.03%
Average	1.29%	-0.21%	-0.29%	-0.30%	-0.48%	0.00%

(a) Random access

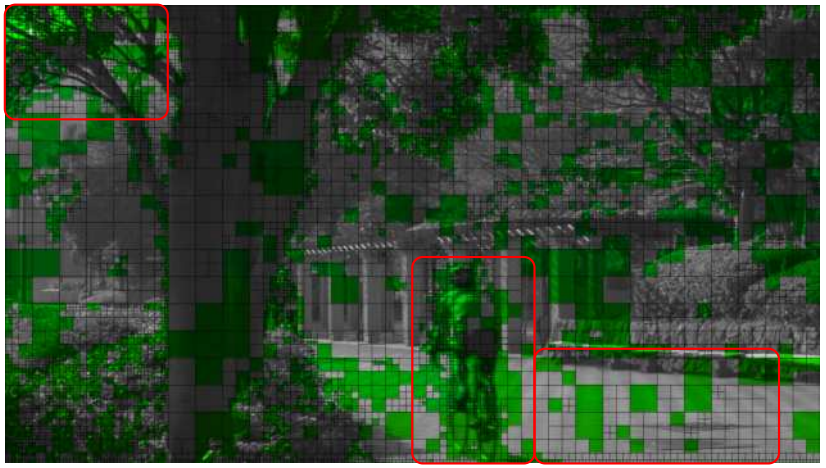
		Low delay P				
Video (1 Mbps)	Sig (%)	4 (%)	8 (%)	16 (%)	32 (%)	64 (%)
Kimono	1.75%	0.03%	-0.12%	1.93%	-3.62%	0.02%
DanceKiss	-1.03%	0.11%	-0.20%	2.48%	-1.34%	-0.03%
FlagShoot	1.74%	0.06%	-0.69%	1.14%	-1.92%	-0.32%
BQTerrace	-4.12%	0.20%	-1.55%	4.81%	0.68%	-0.02%
Average	-0.41%	0.10%	-0.64%	2.59%	-1.55%	-0.09%
Video (16 Mbps)	Sig (%)	4 (%)	8 (%)	16 (%)	32 (%)	64 (%)
Kimono	0.05%	-0.05%	-0.11%	1.42%	-1.32%	0.00%
DanceKiss	-1.37%	0.36%	1.80%	-1.54%	0.71%	0.04%
FlagShoot	-0.94%	0.22%	-0.30%	-0.44%	1.47%	-0.01%
BQTerrace	-0.93%	0.62%	-0.23%	-0.18%	0.72%	0.00%
Average	-0.80%	0.29%	0.29%	-0.18%	0.39%	0.01%

(b) Low delay P

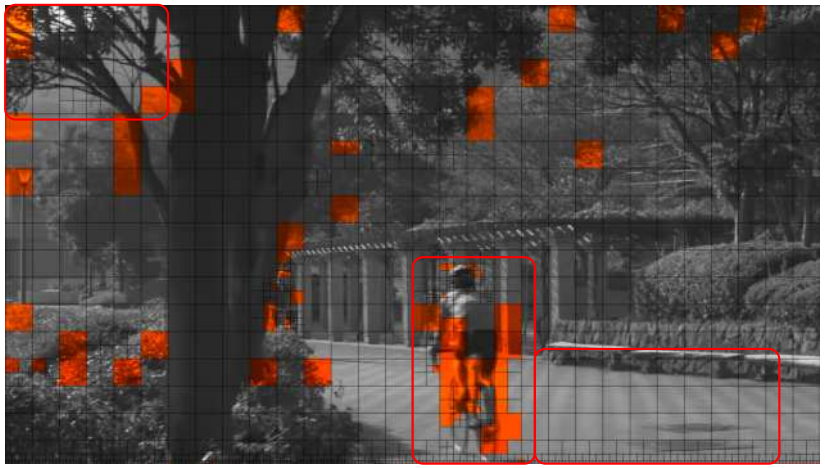
Table 8.1 Changes in bit usage of proposed from original for frame 77 at 1 and 16 Mbps by configuration, for the respective video sequence



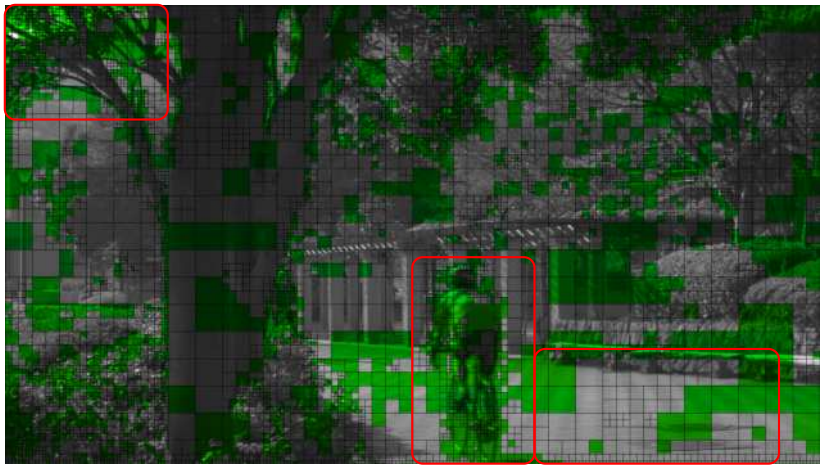
(a) Original 1 Mbps



(b) Original 16 Mbps

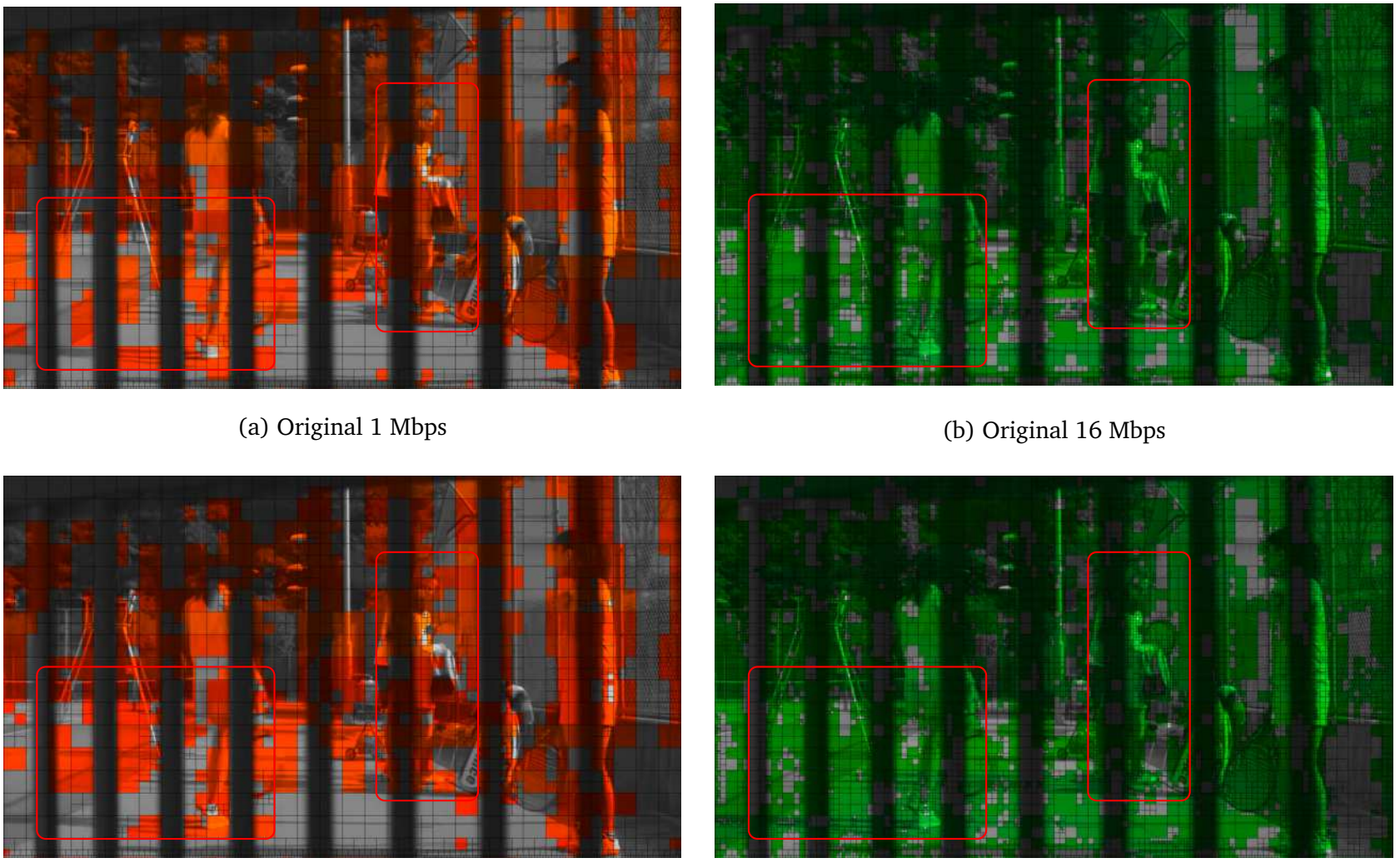


(c) Proposed 1 Mbps



(d) Proposed 16 Mbps

Figure 8.4 Park scene decoded frame 77 with highlighted QP for 1 and 16 Mbps.



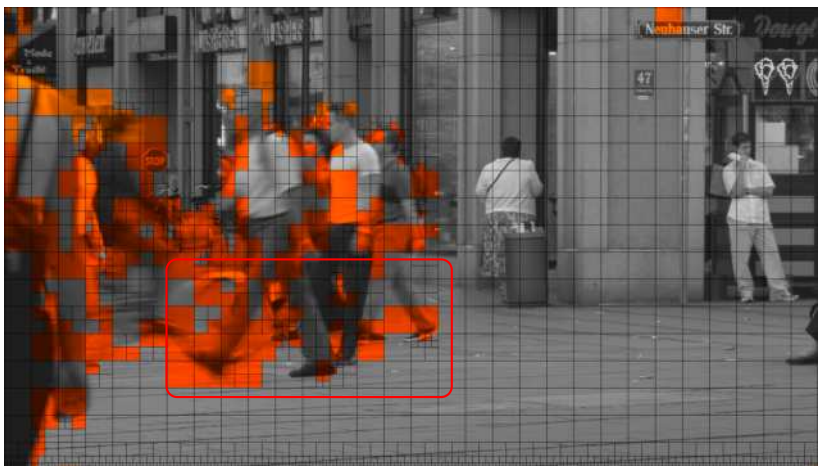
(a) Original 1 Mbps

(b) Original 16 Mbps

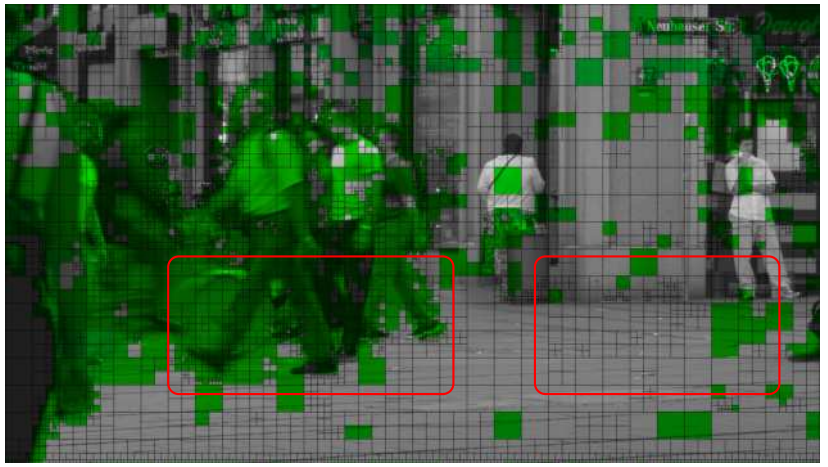
(c) Proposed 1 Mbps

(d) Proposed 16 Mbps

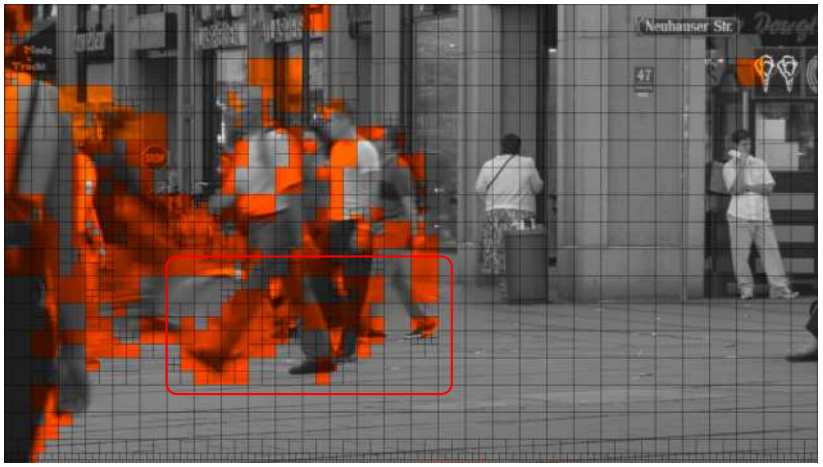
Figure 8.5 Tennis decoded frame 77 with highlighted QP for 1 and 16 Mbps.



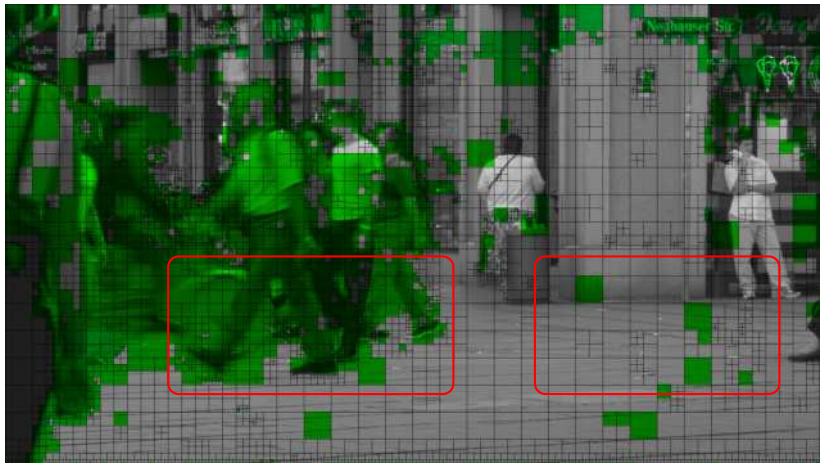
(a) Original 1 Mbps



(b) Original 16 Mbps

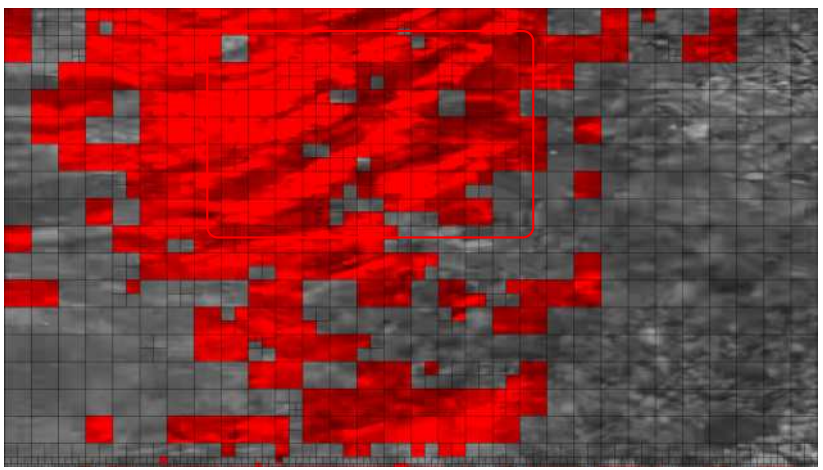


(c) Proposed 1 Mbps

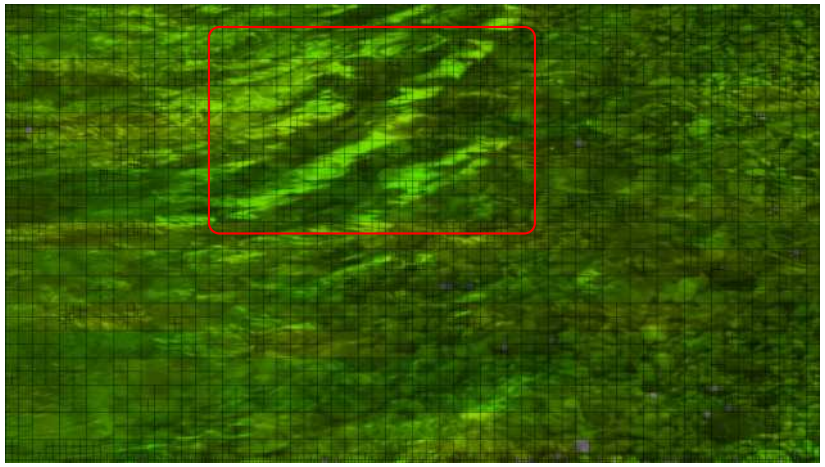


(d) Proposed 16 Mbps

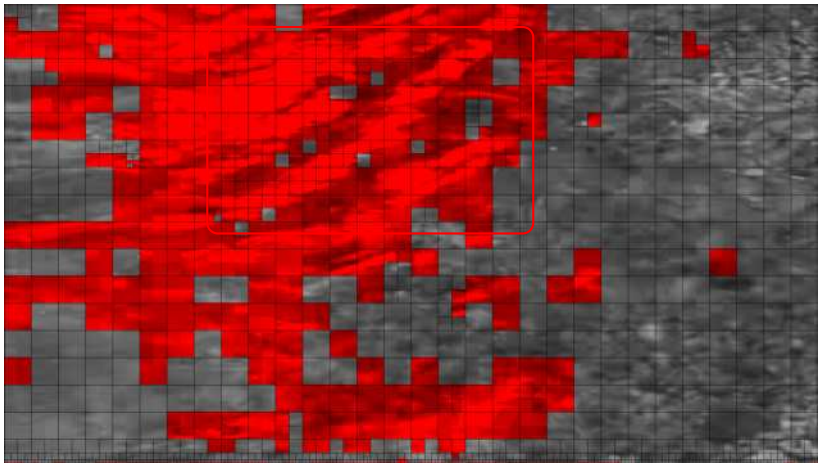
Figure 8.6 Pedestrian area decoded frame 77 with highlighted QP for 1 and 16 Mbps.



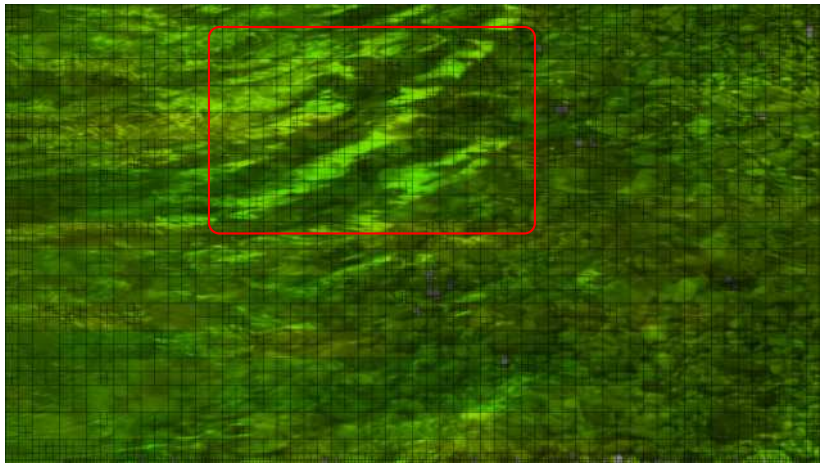
(a) Original 1 Mbps



(b) Original 16 Mbps

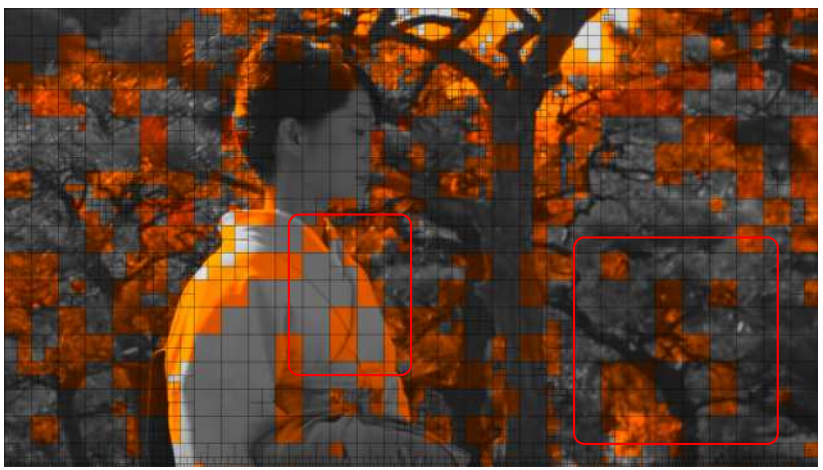


(c) Proposed 1 Mbps



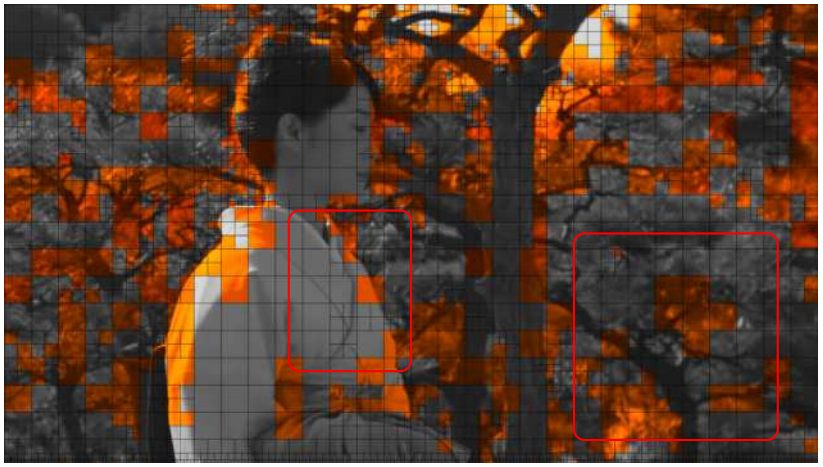
(d) Proposed 16 Mbps

Figure 8.7 Riverbed decoded frame 77 with highlighted QP for 1 and 16 Mbps.

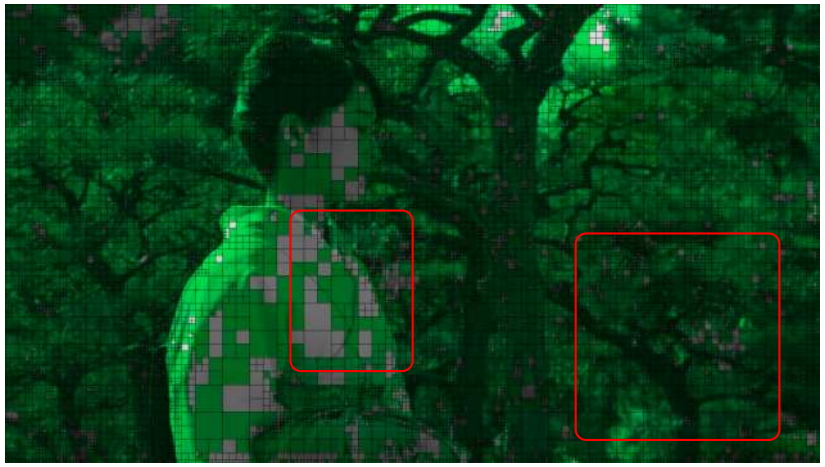


(a) Original 1 Mbps

(b) Original 16 Mbps

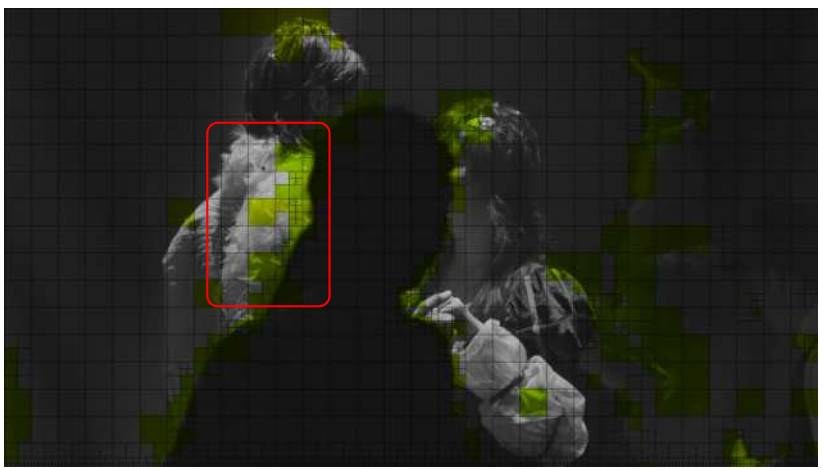


(c) Proposed 1 Mbps

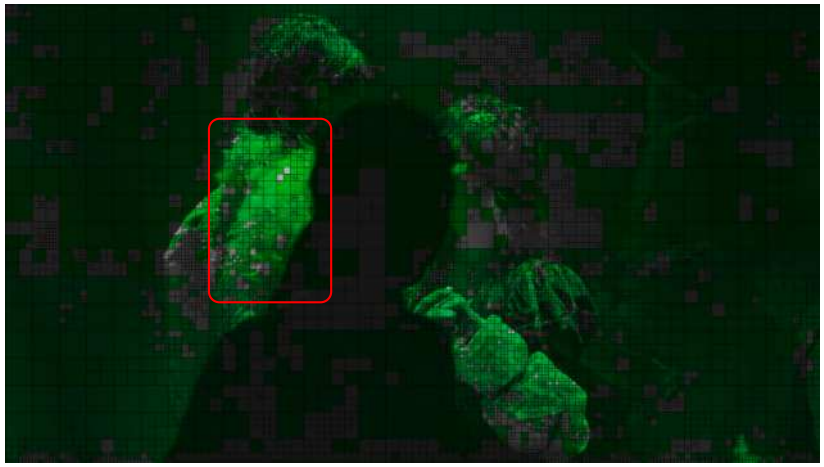


(d) Proposed 16 Mbps

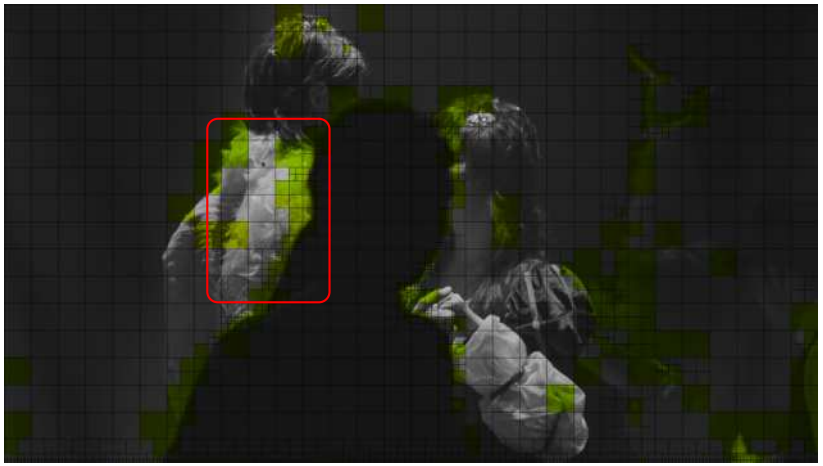
Figure 8.8 Kimono decoded frame 77 with highlighted QP for 1 and 16 Mbps.



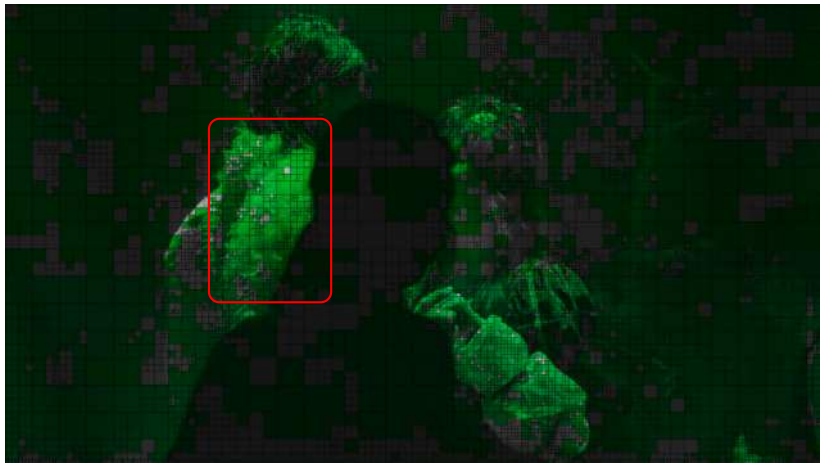
(a) Original 1 Mbps



(b) Original 16 Mbps

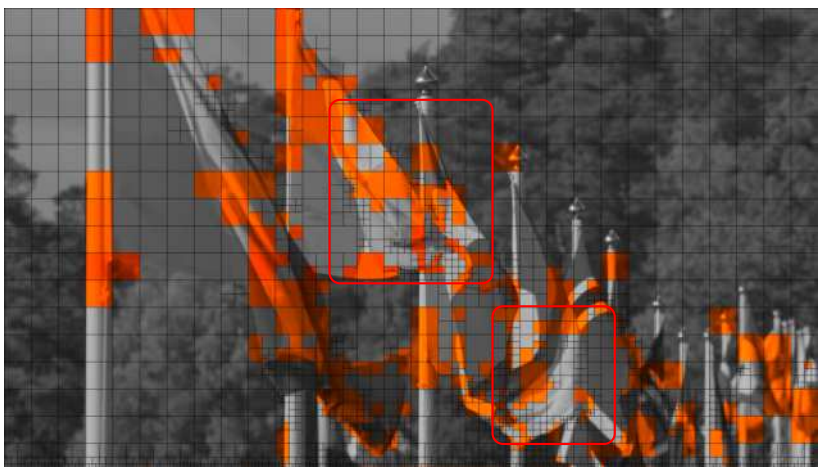


(c) Proposed 1 Mbps

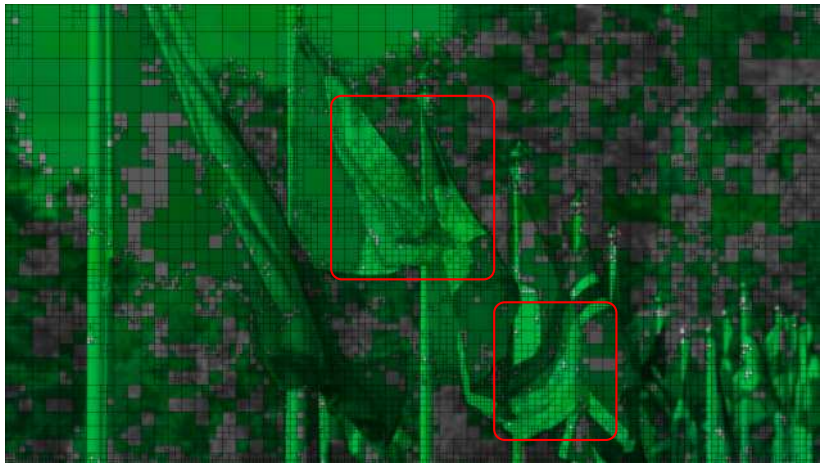


(d) Proposed 16 Mbps

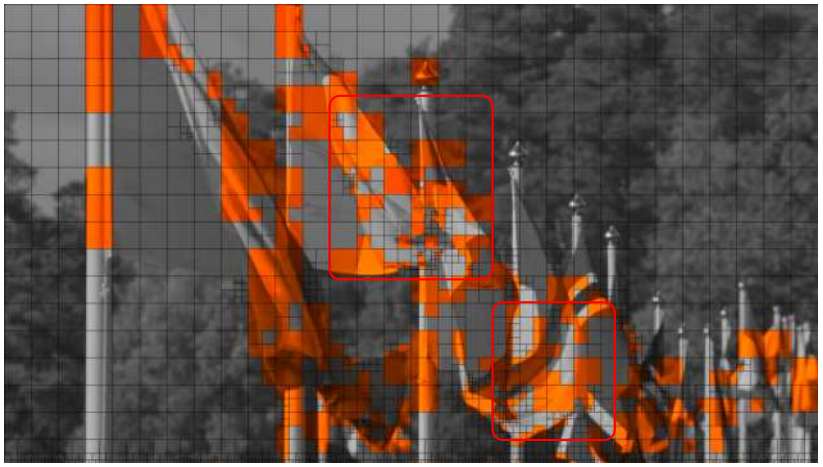
Figure 8.9 DanceKiss decoded frame 77 with highlighted QP for 1 and 16 Mbps.



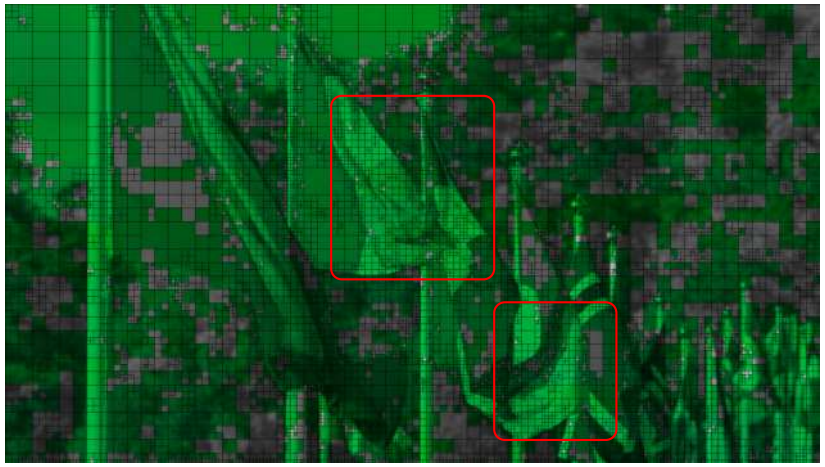
(a) Original 1 Mbps



(b) Original 16 Mbps

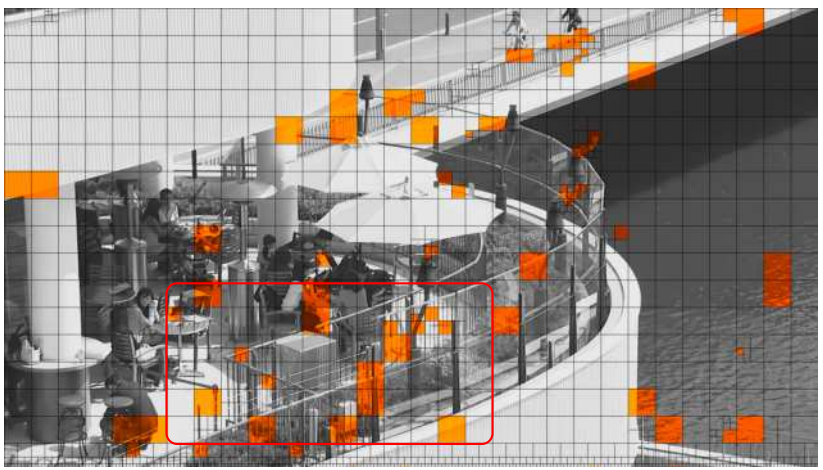


(c) Proposed 1 Mbps

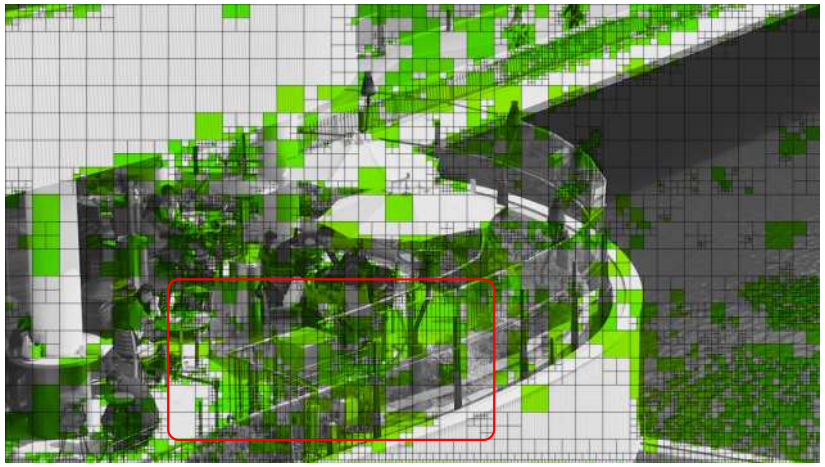


(d) Proposed 16 Mbps

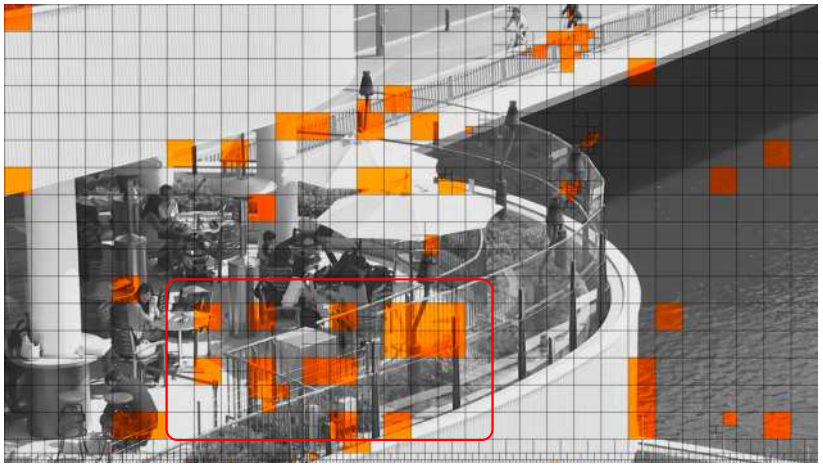
Figure 8.10 FlagShoot decoded frame 77 with highlighted QP for 1 and 16 Mbps.



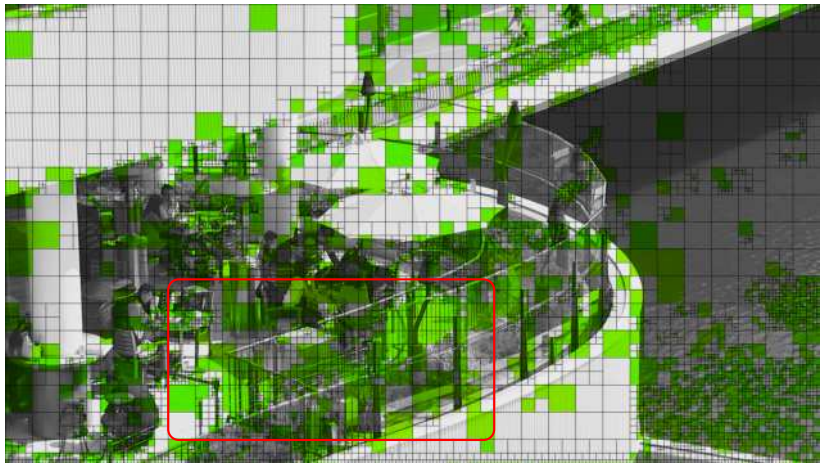
(a) Original 1 Mbps



(b) Original 16 Mbps



(c) Proposed 1 Mbps



(d) Proposed 16 Mbps

Figure 8.11 BQTerrace decoded frame 77 with highlighted QP for 1 and 16 Mbps.

8.5 Visual VCL bit usage distribution

The visual VCL bit usage distribution results are where the LCU bit usage is highlighted as part of a heat map. The tool allows observing where if any bit-redistribution has occurred on the encoded frame. Unlike QP distribution, where scale is bounded between 0 and 51, in bit usage the dynamic range depends upon the content activity and bandwidth allocated. However, for these results the same maximum bits per LCU is used, set at 256. This is applied during Visual VCL heat map generation across both bit-rates of 16 and 1 Mbps. Understandably, for these extreme bit-rates, the heat maps are more likely to have blocks, LCUs which are red at 16Mbps and more blue LCUs in 1Mbps heat maps. Also, the minimum for a heat map colour to be applied is where the LCU bit usage is greater than the maximum bits per LCU divided by 2^8, which represents 0% of the maximum. This equates to where the LCU represents either 0 or 1 bits, in which case the LCU is greyscale, Luma information only. This can be seen in the results shown in Figures 8.16 to 8.19 where the original video sequence frame has the meta signalling and bit usage information superimposed.

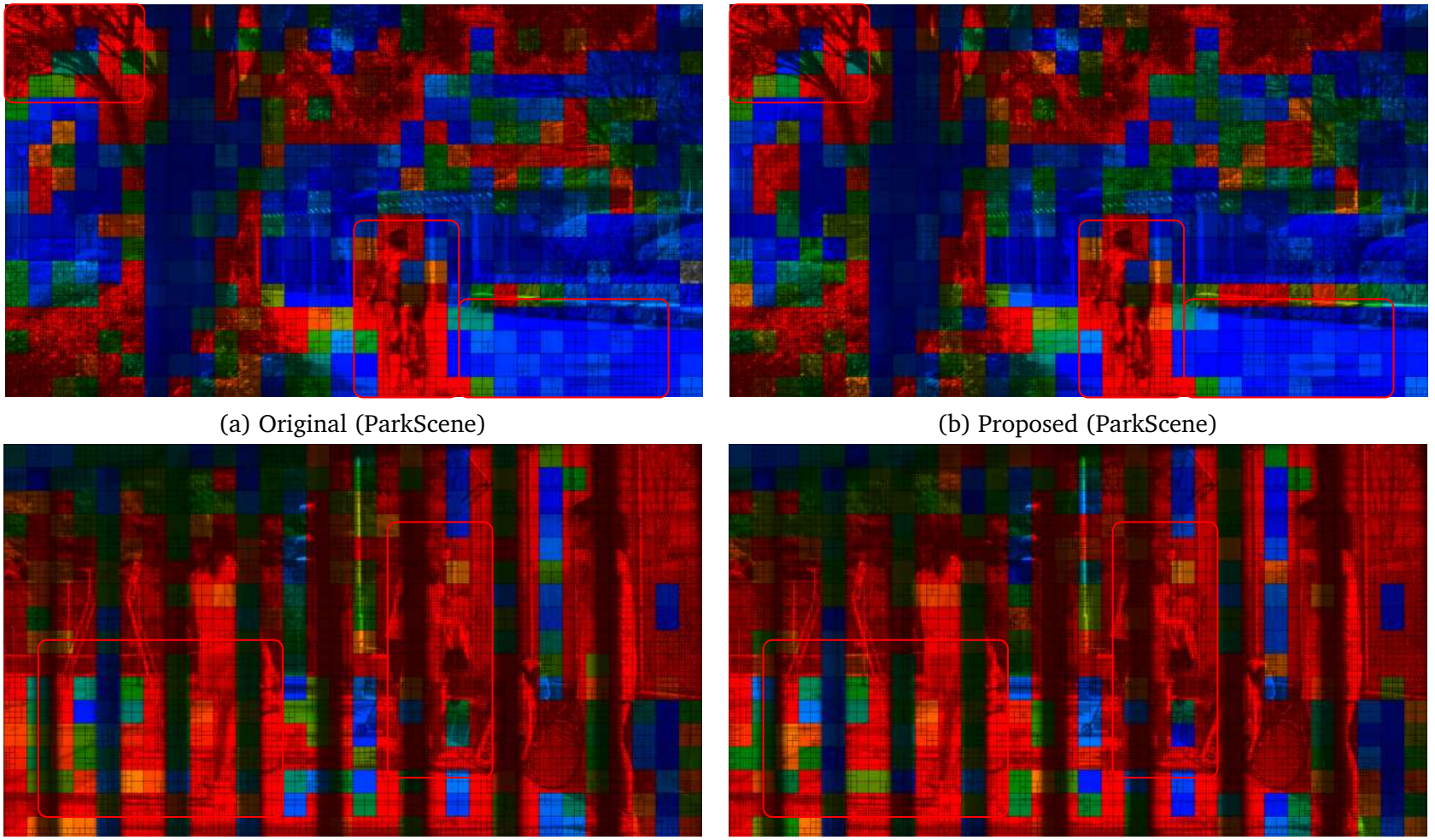
8.5.1 Visual VCL bit usage distribution for frame 77 at 16 Mbps

Using the visual VCL tool to examine the bit distribution for where bandwidth resources are high, such as 16Mbps, can be considered highly difficult to compare, especially where activity is low. In Figures 8.12 to 8.15 the bit usage distribution is illustrated by LCU for the video sequences encoded under random access and low delay P configuration respectively. The differences for random access can be more dynamic compared to low delay P, throughout the changes observed in the proposed encoded video sequences are focused on lowering the bit usage of the perceptually homogeneous regions to encourage bit-redistribution. Using the same encoded video sequences as before, these results can be described by the encoded configuration, first by random access, then low delay P.

For the video sequence ParkScene in Figures 8.12a and 8.12b, observable differences are minor. The proposed encoder, reduces the bit usage in the upper right quadrant around the foliage, while increases the bit usage on the stone brick wall and benches below. Similarly, for Tennis in Figures 8.12c and 8.12d, the

differences occur in the top left corner, where the background texture is being represented more uniformly by the proposed encoder with less high bit usage as it is perceptually homogeneous compared to other regions. In PedestrianArea, activity and texture is more polarised, this means that the heat maps in Figures 8.17a and 8.17b illustrate that the proposed encoder is distinct in the perceptual significant and perceptual homogeneous regions. The proposed encoder is able to redistribute bits for where textures and boundaries occur. When the respective encoders are tested with Riverbed in Figures 8.17c and 8.17d, the high level of localised activity means the bit usage coverage is similar.

For low delay P, the changes occur in similar types of regions, yet the effects are more subtle. In these results, the shifts in heat map banding of colours are difficult to observe and instead the form, shape or direction of these changes must be considered. This can be shown in Figures 8.14a and 8.14b, where the net difference suggest they cancel out, however, the arrangement of blue LCUs in the proposed encoder suggest a perceptual homogeneous region across the video frame. When the proposed encoder is applied to DanceKiss, a video sequence taken indoors, the results in Figures 8.14c and 8.14d shows the proposed encoder is able to lower the LCU bit usage on left and right hand side of the video frame, where the background is perceptually homogeneous. For an outdoor active video sequence, like FlagShoot, as Figures 8.14c and 8.14d shows, the changes are related to retain perceptual integrity, by assigning more bits per LCU to capture the background, where the trees are. From the Visual VCL bit usage, BQTerrace is quite difficult challenging for an encoder, perhaps due to the high frame rate of 60fps and mixed activity. This is because these set of results, Figures 8.15c and 8.15d are the only ones where despite having the 16Mbps the encoder chooses to allocate LCU with either one or no bits. The proposed encoder continues its trend and is able to offer a lower bit usage on the upper left hand quadrant of the frame and provide more bits per LCU along the terrace glass frontage.



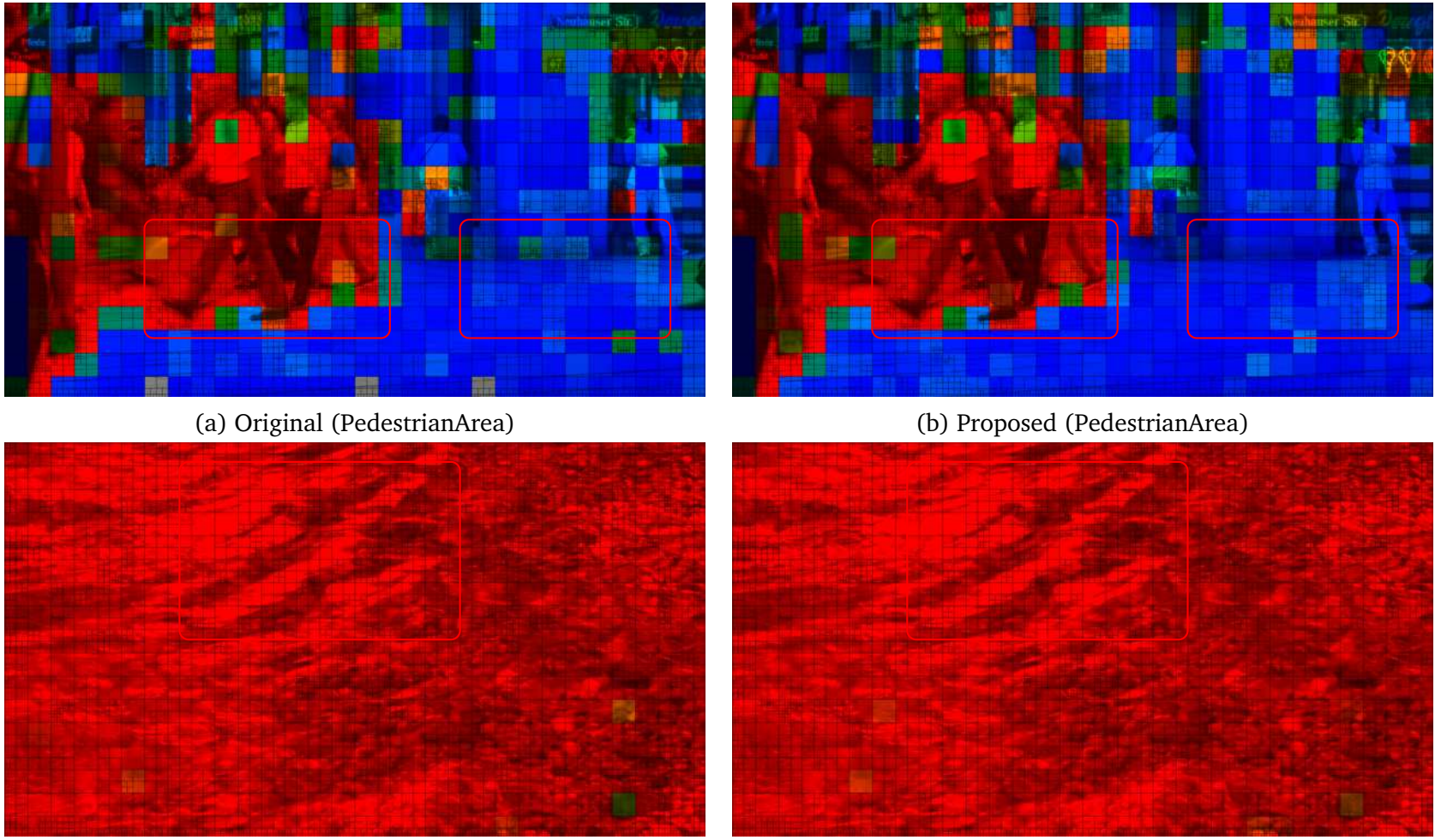
(a) Original (ParkScene)

(b) Proposed (ParkScene)

(c) Original (Tennis)

(d) Proposed (Tennis)

Figure 8.12 Bit usage by LCU, frame 77 for random access encoded sequences ParkScene and Tennis at 16Mbps, where maximum bit per coding unit is 256



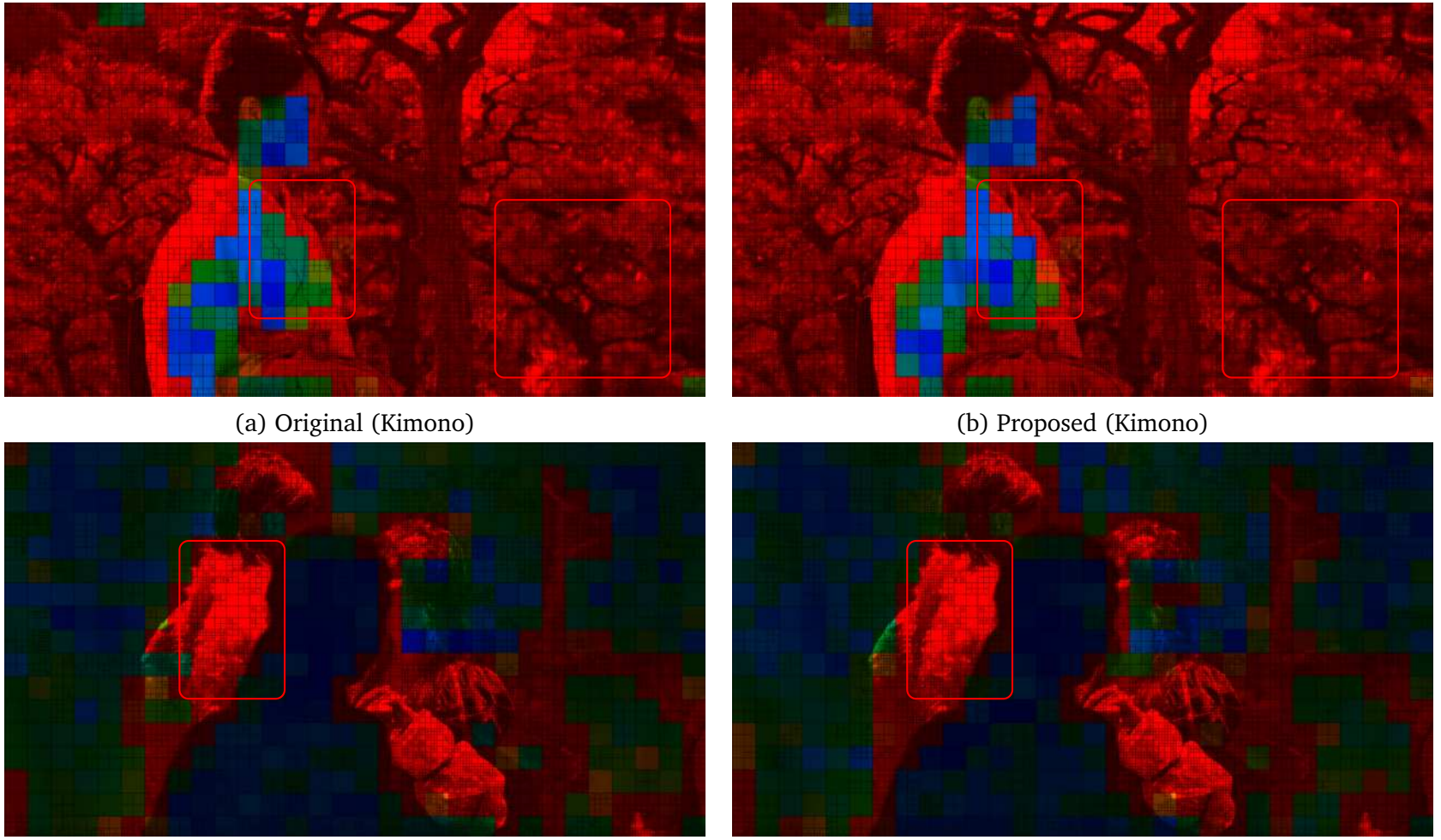
(a) Original (PedestrianArea)

(b) Proposed (PedestrianArea)

(c) Original (Riverbed)

(d) Proposed (Riverbed)

Figure 8.13 Bit usage by LCU, frame 77 for random access encoded sequences PedestrianArea and Riverbed at 16Mbps, where maximum bit per coding unit is 256



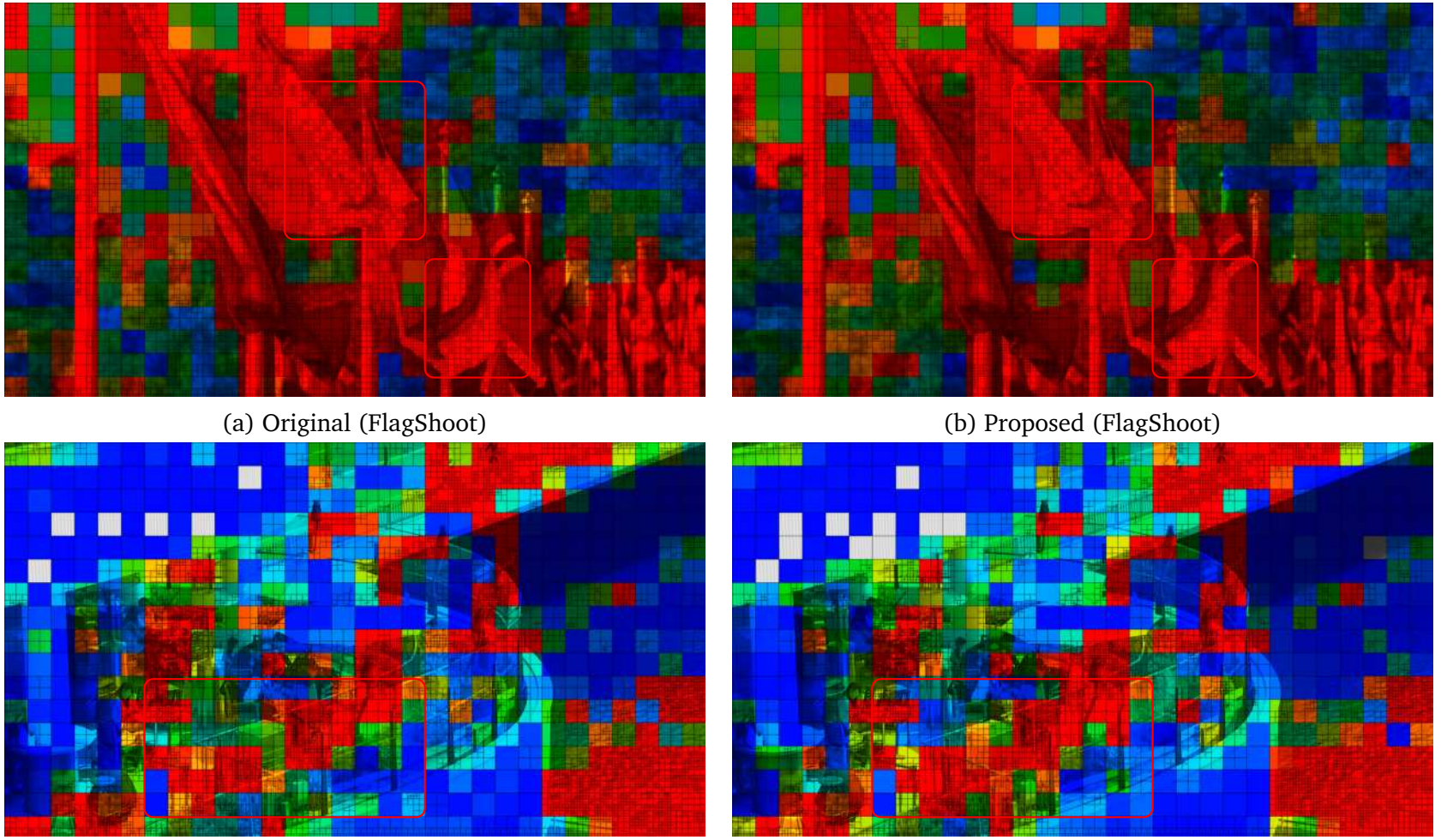
(a) Original (Kimono)

(b) Proposed (Kimono)

(c) Original (DanceKiss)

(d) Proposed (DanceKiss)

Figure 8.14 Bit usage by LCU, frame 77 for low delay P encoded sequences Kimono and DanceKiss at 16Mbps, where maximum bits per coding unit is 256



(a) Original (FlagShoot)

(b) Proposed (FlagShoot)

(c) Original (BQTerrace)

(d) Proposed (BQTerrace)

Figure 8.15 Bit usage by LCU, frame 77 for low delay P encoded sequences FlagShoot and BQTerrace at 16Mbps, where maximum bits per coding unit is 256

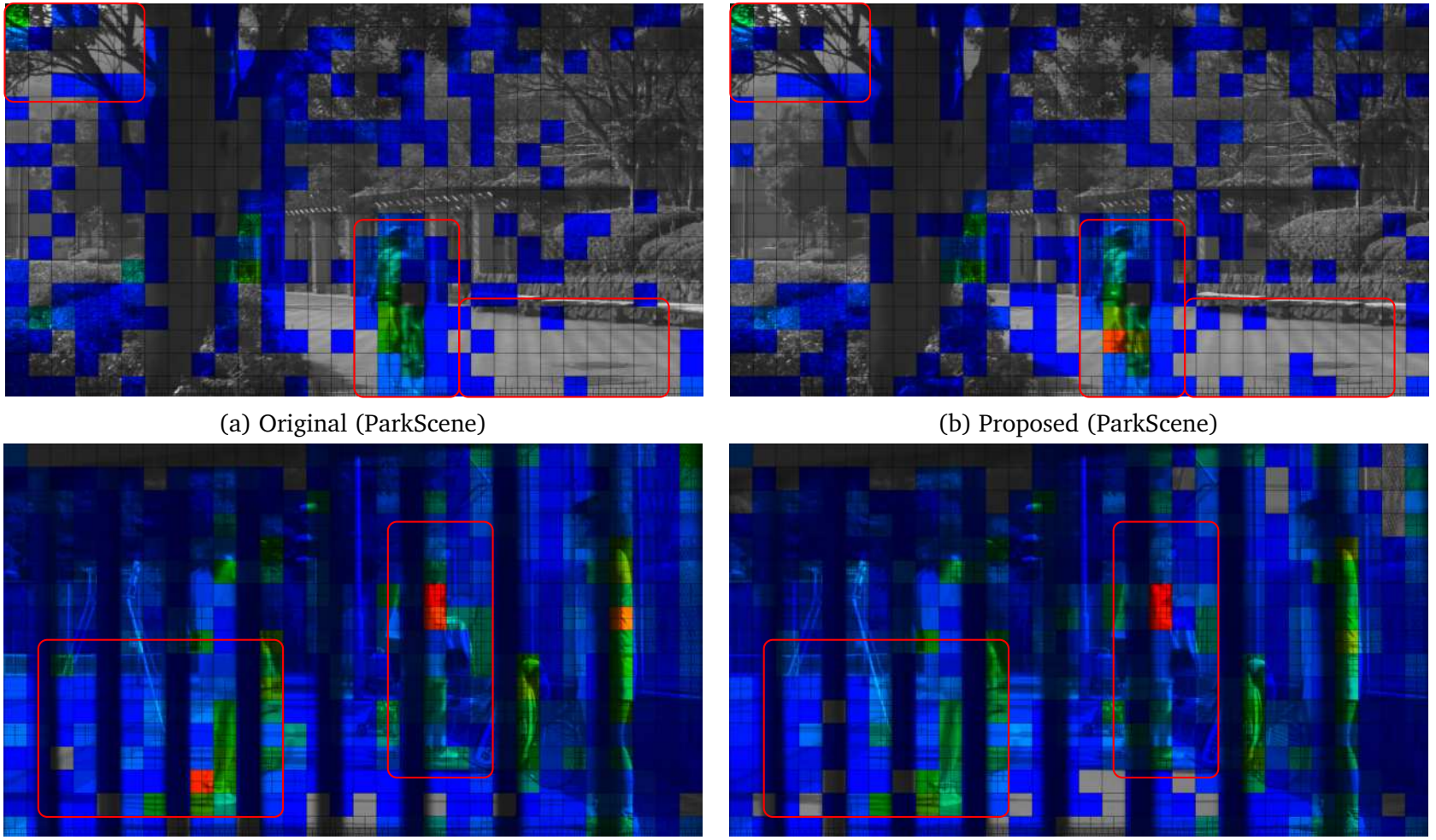
Visual VCL bit usage distribution for frame 77 at 1 Mbps

The 1Mbps Visual VCL bit usage results represent the low extreme end of video coding with the results shown in Figures 8.16 to 8.19 for random access and low delay P configurations. The proposed encoder through these results show that bit-redistribution by identifying perceptually homogeneous regions and representing it with fewer bits. The random access set of results in Figures 8.16 and 8.17 compared to low delay P in Figures 8.18 and 8.19 can demonstrate that the proposed encoder can identify perceptually homogeneous regions. This builds upon the findings seen in the 16Mbps Visual VCL bit usage results, where perceptually homogeneous regions were shown to be allocated fewer bits per LCU.

In the random access configuration result of ParkScene, in Figures 8.16a and 8.16b, the proposed encoder reduces bit usage on the left hand side of the frame, which encourages more bits on and around the cyclist. For Tennis, the results shown in Figures 8.16c and 8.16d, highlight the proposed is able to provide more LCUs with one or zero bits. Furthermore, the peak bit usage is isolated to individuals at the centre of the frame, providing a uniform bit spread elsewhere. The proposed encoder is able to reduce bit usage of spurious non-perceptual LCUs particularly well in PedrestrianArea, as shown in Figures 8.17a and 8.17b, where on the right hand side the proposed encoder has more LCUs containing one or no bits. This behaviour is repeated with Riverbed in Figures 8.17c and 8.17d, where with the proposed encoder both the right hand side and bottom left quadrant illustrate more LCUs containing one or zero bits, demonstrating that perceptual bit-redistribution is occurring.

For the results in low delay P, the changes between encoders are more subtle, with similar theme of increasing the spread of bit usage to cover the largest area of perceptual significance. This could mean fewer zero or one bit LCUs if the textures are perceptually significant, such as that shown with Kimono in Figures 8.18a and 8.18b, where the proposed tries to cover the background textures of the foliage. While for DanceKiss in Figures 8.18c and 8.18d the proposed encoder is able increase the LCUs with the least bit usage on the left hand side of the frame. For FlagShoot, the results shown in Figures 8.19a and 8.19b present a similar bit usage distribution, with differences occurring outside the diagonal flags from the

top left to the bottom right. In BQTerrace where widespread activity occurs, the proposed encoder provides more LCUs with bit usage greater than 1 bit minimum, suggesting that the proposed encoder has identified perceptually significant regions and is spreading the bits across a wider number of LCUs.



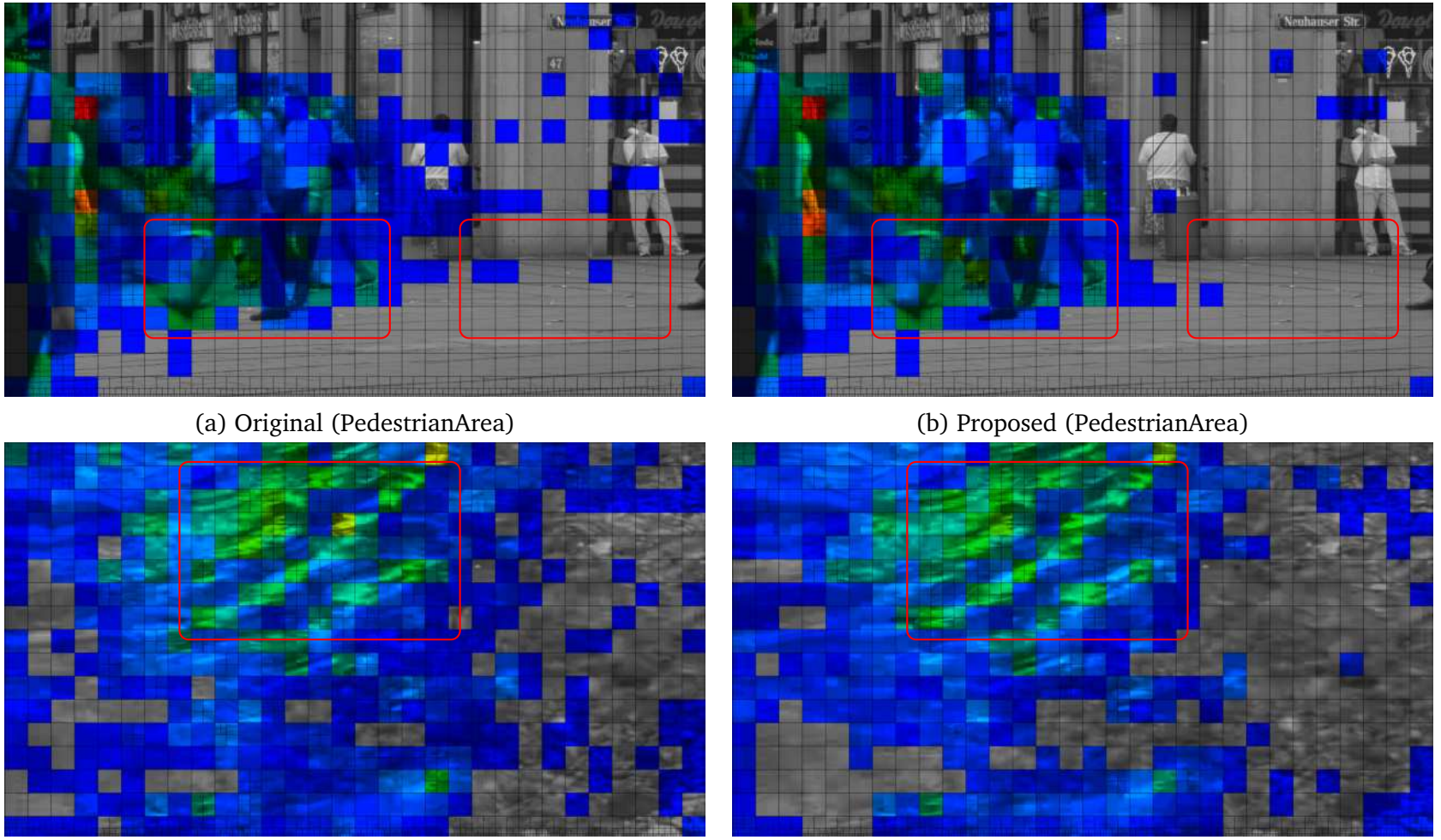
(a) Original (ParkScene)

(b) Proposed (ParkScene)

(c) Original (Tennis)

(d) Proposed (Tennis)

Figure 8.16 Bit usage by LCU, frame 77 for random access encoded sequences ParkScene and Tennis at 1Mbps, where maximum bit per coding unit is 256



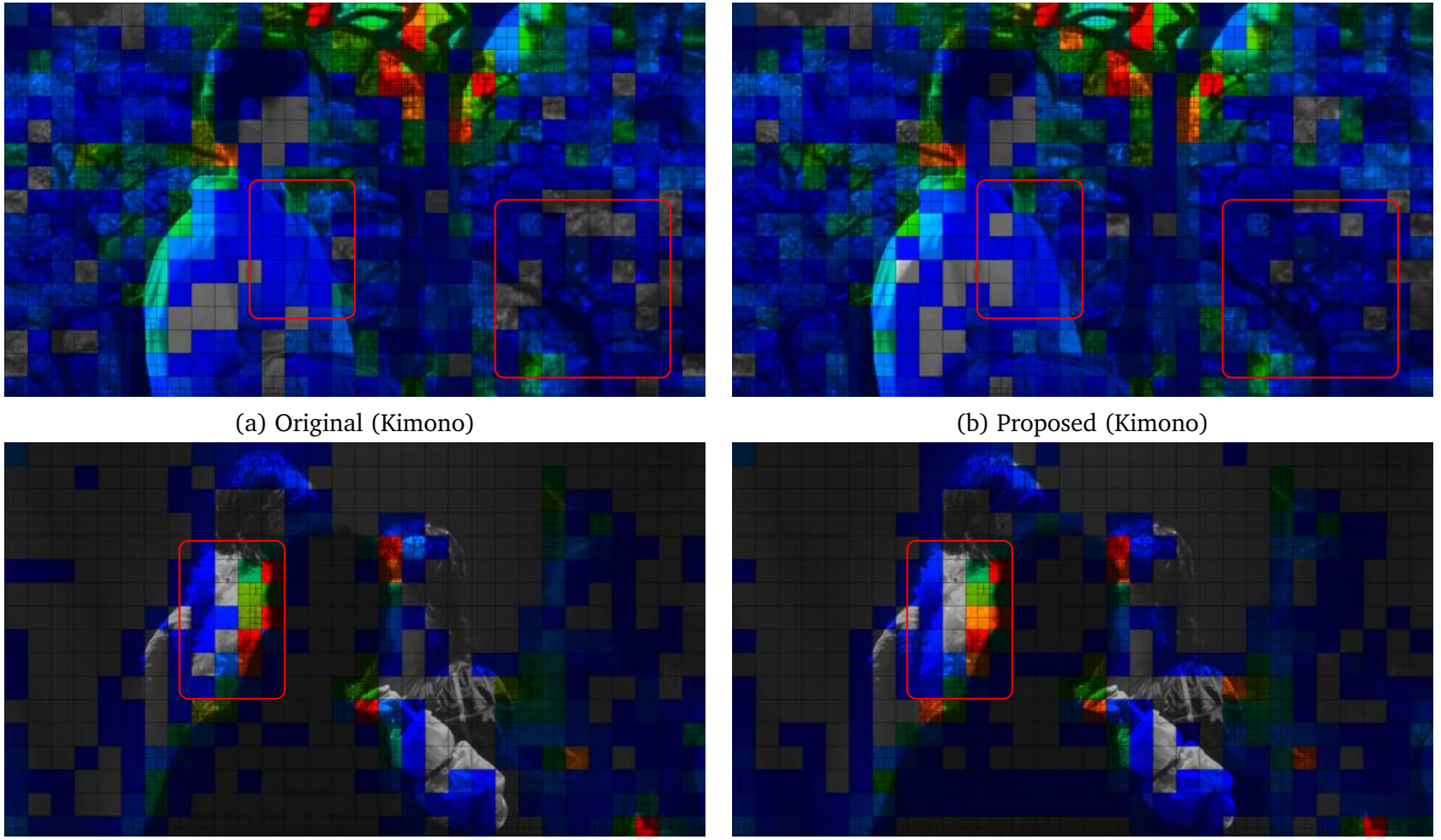
(a) Original (PedestrianArea)

(b) Proposed (PedestrianArea)

(c) Original (Riverbed)

(d) Proposed (Riverbed)

Figure 8.17 Bit usage by LCU, frame 77 for random access encoded sequences PedestrianArea and Riverbed at 1Mbps, where maximum bit per coding unit is 256



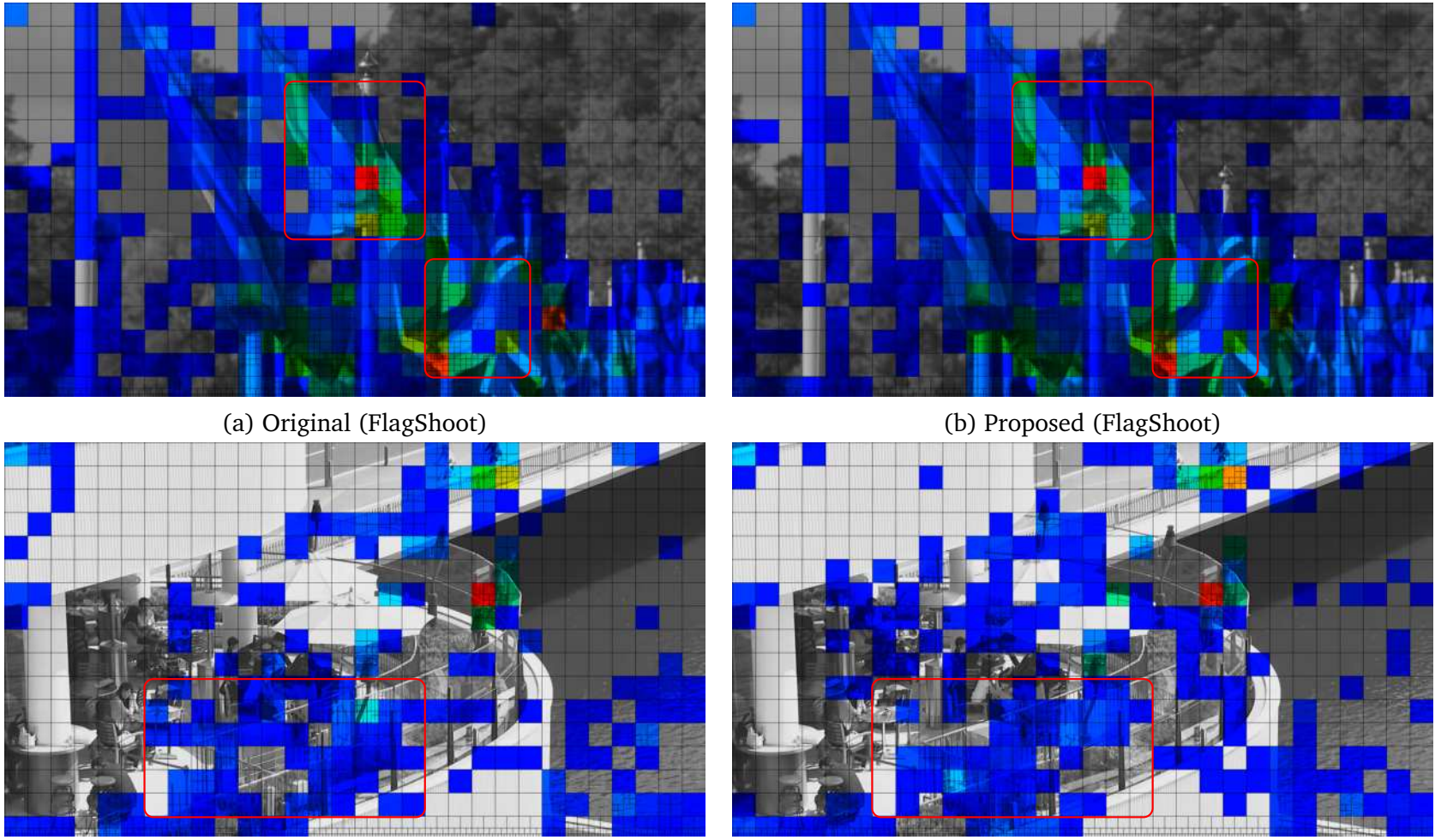
(a) Original (Kimono)

(b) Proposed (Kimono)

(c) Original (DanceKiss)

(d) Proposed (DanceKiss)

Figure 8.18 Bit usage by LCU, frame 77 for low delay P encoded sequences Kimono and DanceKiss at 1Mbps, where maximum bits per coding unit is 256



(a) Original (FlagShoot)

(b) Proposed (FlagShoot)

(c) Original (BQTerrace)

(d) Proposed (BQTerrace)

Figure 8.19 Bit usage by LCU, frame 77 for low delay P encoded sequences FlagShoot and BQTerrace at 1Mbps, where maximum bits per coding unit is 256

8.6 Visual VCL assessment

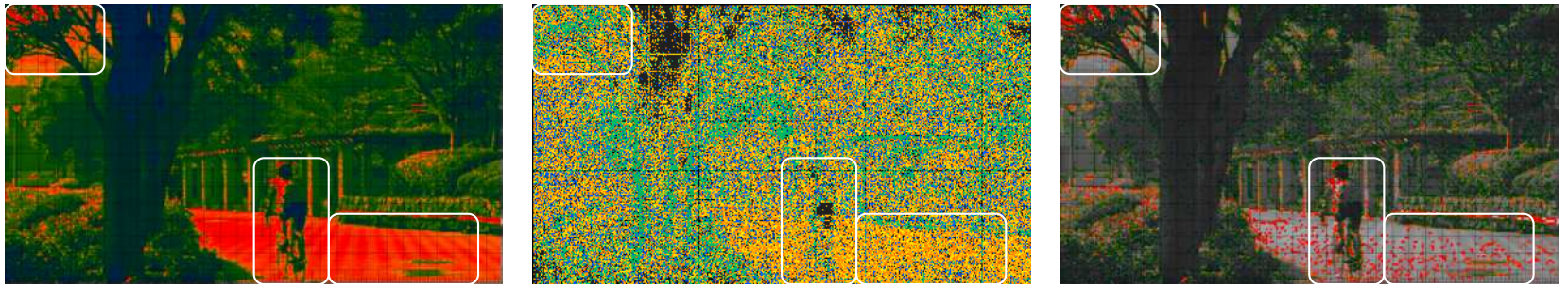
Another feature of the Visual VCL Tool is the ability to simulate assessment, which can be used on the video sequences to visualise rate-control activity and the perceptual effect of the respective encoders. Being able to visual the rate-control can demonstrate who the respective encoders would allocate bit-budgets per LCU. While the ability to visualise the perceptual assessment, can demonstrate which encoder was able to retain perceptual clues. For rate-control, activity assessment is based upon the original video frame, irrespective of the encoded bitstream and this is shown as triplets of existing rate-control Hadamard, JND and the proposed ppwAPC in Figures 8.20 to 8.27. Perceptual assessment is provided by SSIM and is applied at 1 Mbps where distortion is most apparent compared to the other higher bit-rates. In all, visualising the VCL using these assessments can assist to understand the effect of choices in allocating bits.

8.6.1 Rate-control

The rate-control activity covers three sets of assessment, the existing reference of Hadamard 8x8, a perceptual model of JND and the proposed simulated Hadamard 8x8 with ppwAPC. To allow a comparison to be shown together the results are shown as triplets in Figures 8.20 to 8.27, however, each have some variation in their colour scheme. The Hadamard based activity is illustrated in monochrome or tri-colour manner of red, green and black, with red being the highest score. JND renders the scene as one of five colours of yellow (highest), orange, green, blue and black (lowest). While for the proposed ppwAPC, this is applied conditionally, if conditions are met, then it is enabled with a colour of red (highest) or green (lowest).

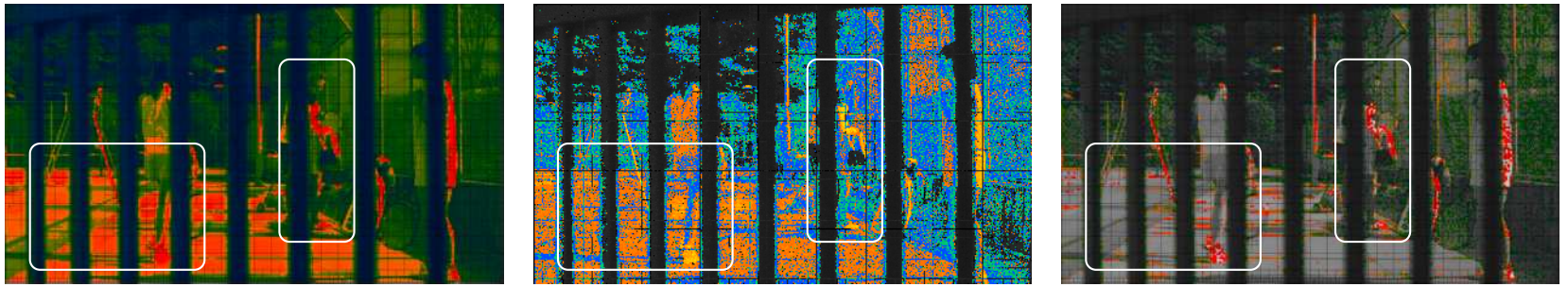
In all a red is the highest value followed by orange or yellow, and it is applied to areas with bright regions and black to shadows, in particular ppwAPC limits itself to where sudden changes in brightness occur. This is shown where bright textured regions exist or where boundaries for well lit object occur such as Tennis and FlagShoot Figures 8.21c and 8.26c. In some frames the activity by ppwAPC is dense in regions as shown in Figures 8.23 and 8.25 to 8.27, which occurs on very bright regions. These high intense areas occur where there are reflections on

the ripples in Riverbed, the spot light on the feather scarf and sleeve of the dress in DanceKiss and the bright walls in BQTerrace. This intensity is constant in each of those triplets, however, in other frames, in Figures 8.20 to 8.22, 8.24 and 8.26 where Hadamard and JND may allocated high scores, ppwAPC is not triggered. This is because these areas are bright, yet their texture are not of perceptual interest, do not trigger ppwAPC, this is why ppwAPC triggers where there is boundary changes to minimise undue additional cost for activity assessment. Finally, as ppwAPC applies perceptual significant test via the fixed sub-block size of 8x8 under rate-control, for which the simulation is a fair reflection of how it operates within the encoder.



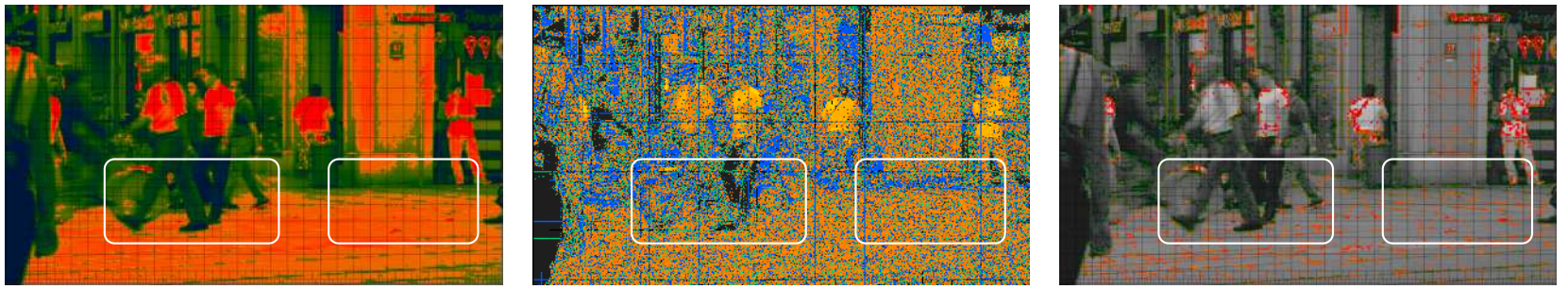
(a) Rate-control Hadamard 8x8 (b) JND (c) ppwAPC

Figure 8.20 Video sequence ParkScene frame 77, simulated rate-control Hadamard 8x8, JND and ppwAPC



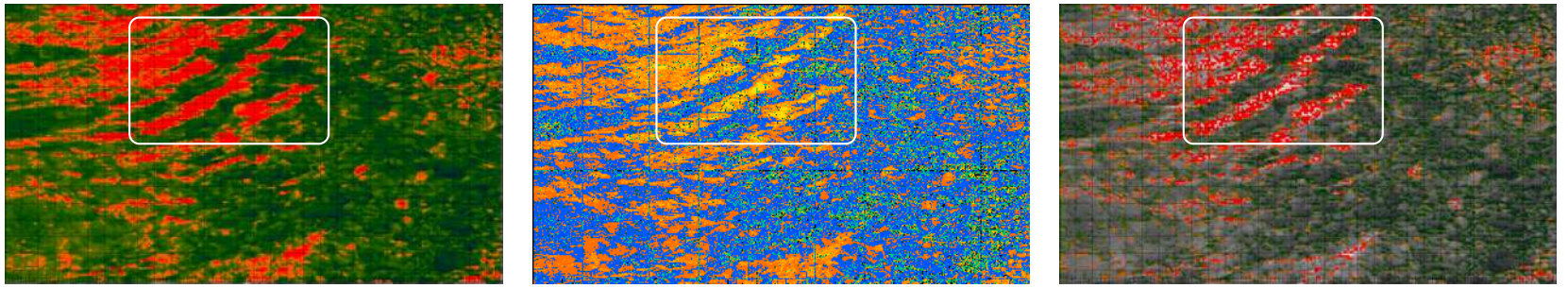
(a) Rate-control Hadamard 8x8 (b) JND (c) ppwAPC

Figure 8.21 Video sequence Tennis frame 77, simulated rate-control Hadamard 8x8, JND and ppwAPC



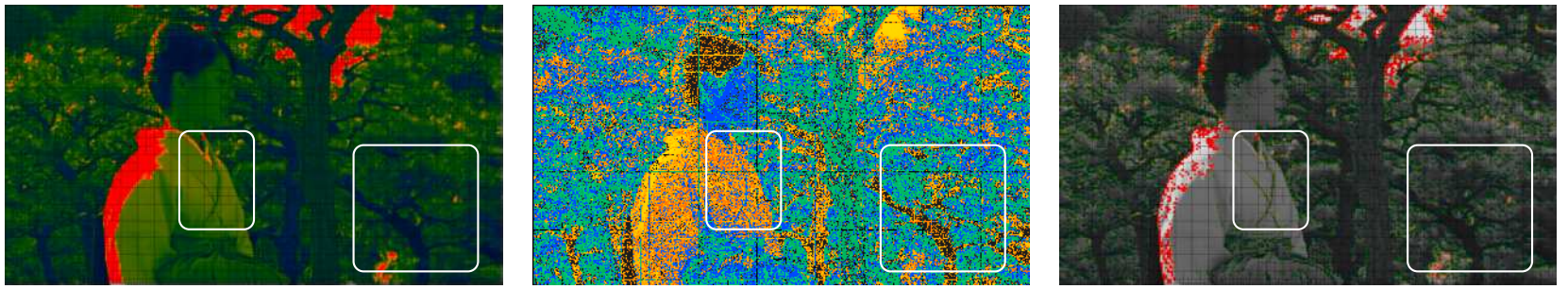
(a) Rate-control Hadamard 8x8 (b) JND (c) ppwAPC

Figure 8.22 Video sequence PedestrianArea frame 77, simulated rate-control Hadamard 8x8, JND and ppwAPC



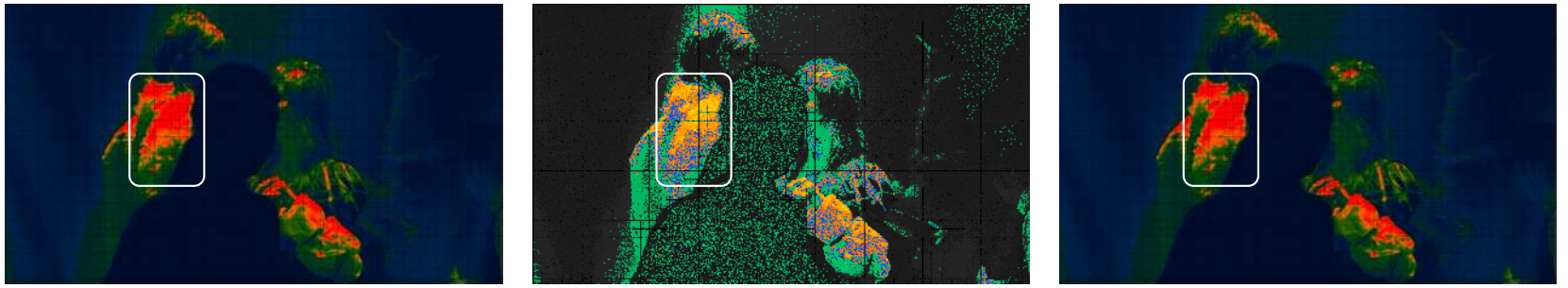
(a) Rate-control Hadamard 8x8 (b) JND (c) ppwAPC

Figure 8.23 Video sequence Riverbed frame 77, simulated rate-control Hadamard 8x8, JND and ppwAPC



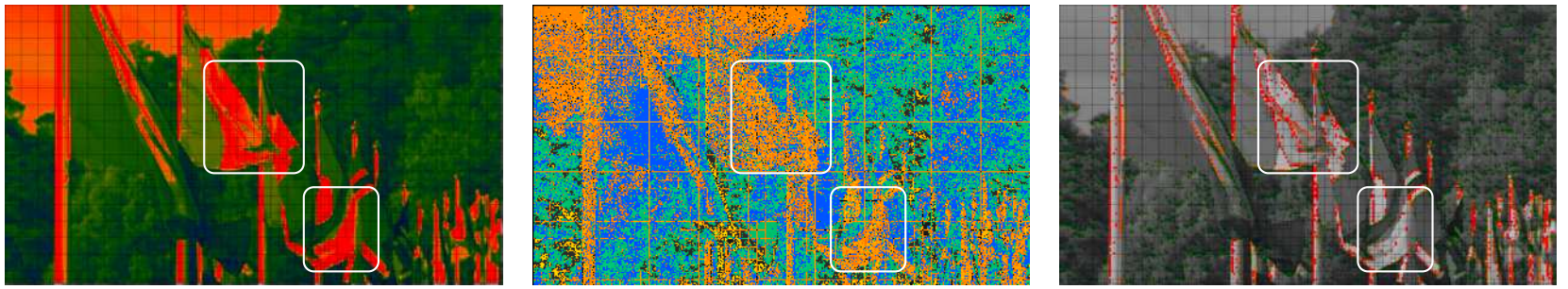
(a) Rate-control Hadamard 8x8 (b) JND (c) ppwAPC

Figure 8.24 Video sequence Kimono frame 77, simulated rate-control Hadamard 8x8, JND and ppwAPC



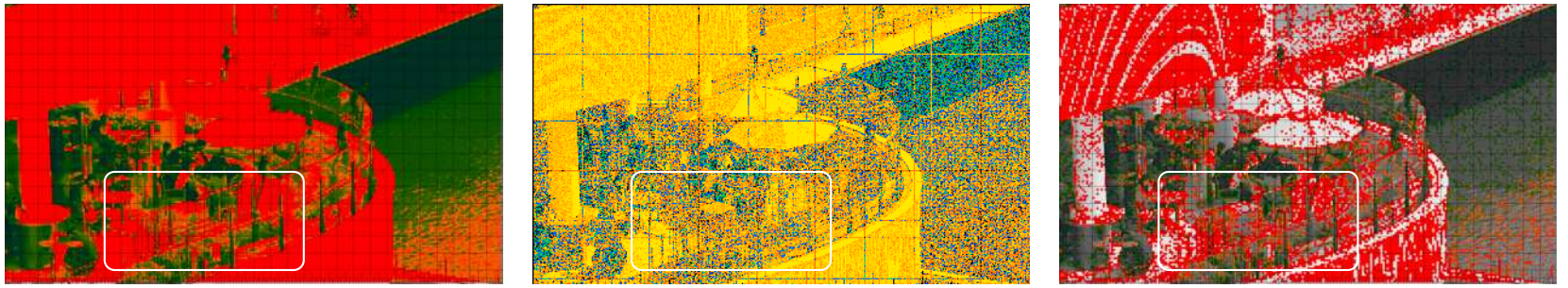
(a) Rate-control Hadamard 8x8 (b) JND (c) ppwAPC

Figure 8.25 Video sequence DanceKiss frame 77, simulated rate-control Hadamard 8x8, JND and ppwAPC



(a) Rate-control Hadamard 8x8 (b) JND (c) ppwAPC

Figure 8.26 Video sequence FlagShoot frame 77, simulated rate-control Hadamard 8x8, JND and ppwAPC

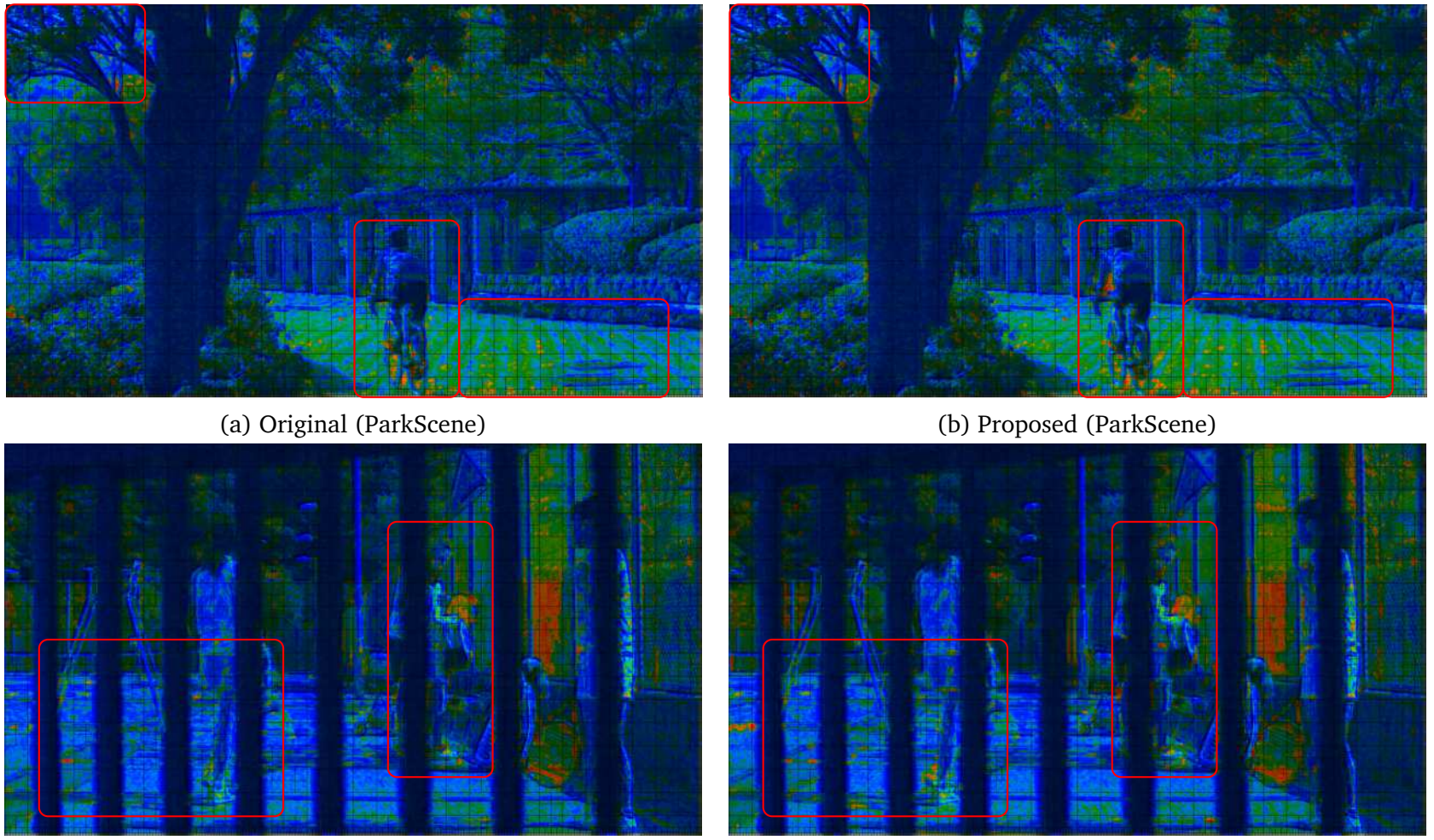


(a) Rate-control Hadamard 8x8 (b) JND (c) ppwAPC

Figure 8.27 Video sequence BQTerrace frame 77, simulated rate-control Hadamard 8x8, JND and ppwAPC

8.6.2 SSIM

SSIM is recognised as a credible means for evaluating perceptual assessment and since SSIM uses a fixed scale of ± 1, this makes it convenient to visualise. As distortion is most present where bit-rate is least, this meant that frames from 1Mbps encoded sequences were used. Overall, the SSIM heat maps in Figures 8.28 to 8.31 illustrate that the amount of perceptual related distortion across the frame is highly dependent upon the nature of the video sequence. When examining ParkScene and Tennis in Figures 8.28a to 8.28d, larger continuous regions of distortion by way of red spots are present in the proposed compared to the original. An example of this is shown on the left arm and body of the cyclist, or at the right foot of the far right tennis player. This behaviour is less likely in less bright regions, or in highly active scenes such as PedestrianArea in Figures 8.29a and 8.29b or Riverbed in Figures 8.29c and 8.29d respectively. From the PedestrianArea SSIM frame, this could be due to relatively low contrast in the video source which is deemed perceptually unimportant. While for Riverbed, the high activity is a struggle for the encoder. The SSIM heat map frames for low delay P in Figures 8.30 and 8.31 show virtually no difference despite both Kimono and BQTerrace, Figures 8.30a and 8.30b, Figures 8.31c and 8.31d respectively, having a high dynamic range. Overall, The proposed encoder is unable to show changes in visually using SSIM assessment for low delay P configuration.



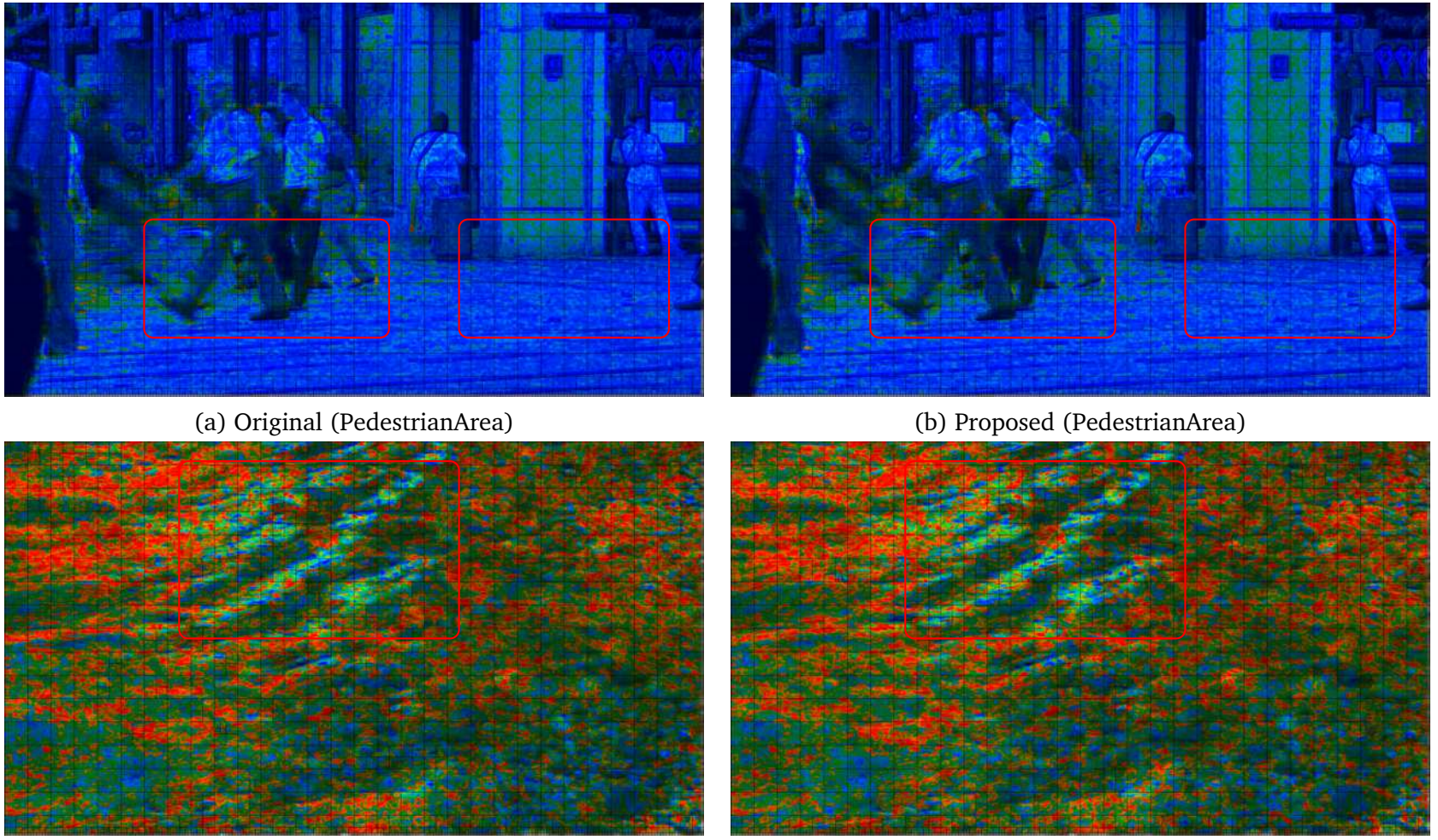
(a) Original (ParkScene)

(b) Proposed (ParkScene)

(c) Original (Tennis)

(d) Proposed (Tennis)

Figure 8.28 Frame 77 SSIM heat map for random access encoded sequences ParkScene and Tennis at 1Mbps



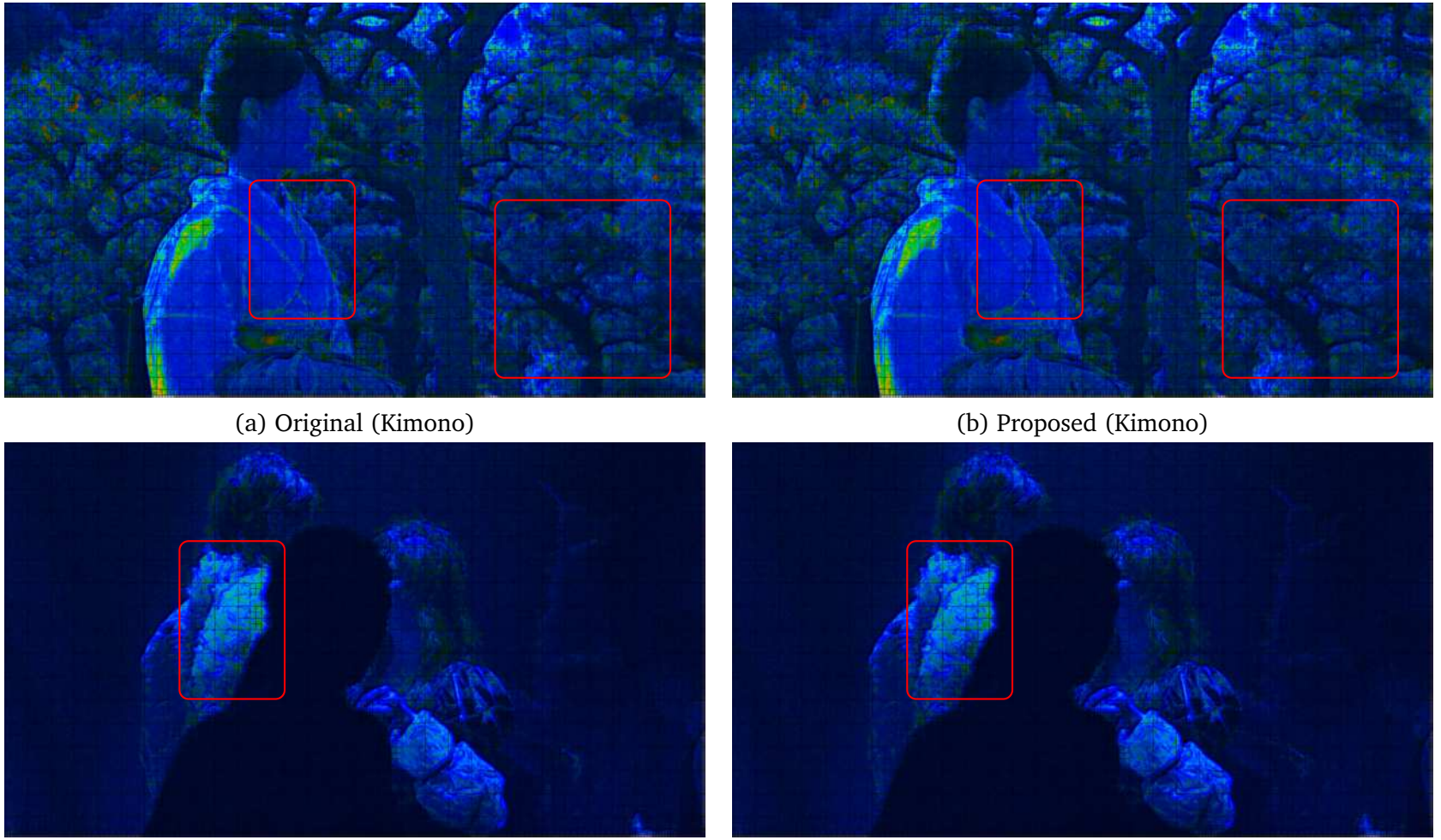
(a) Original (PedestrianArea)

(b) Proposed (PedestrianArea)

(c) Original (Riverbed)

(d) Proposed (Riverbed)

Figure 8.29 Frame 77 SSIM heat map for random access encoded sequences PedestrianArea and Riverbed at 1Mbps



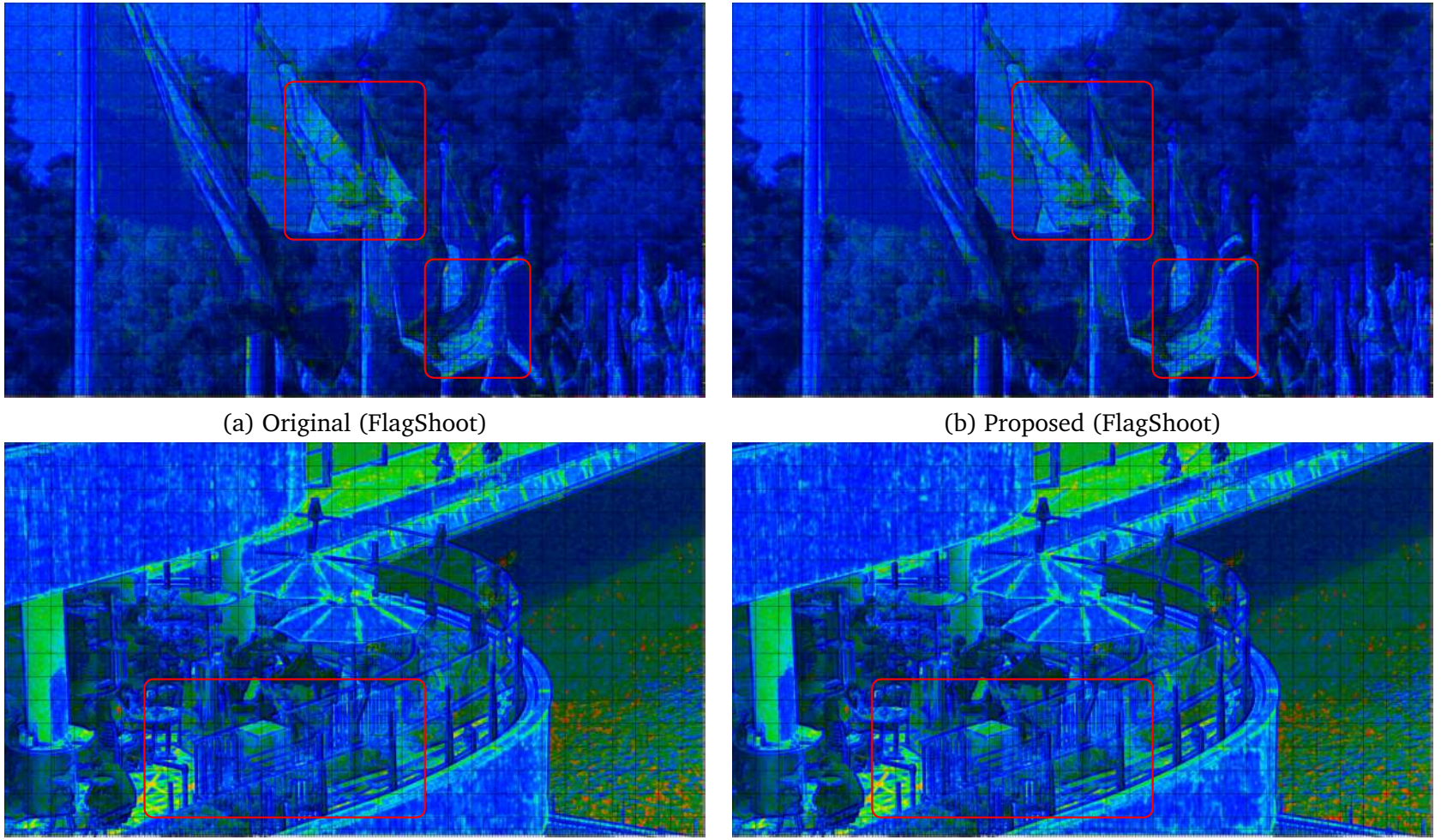
(a) Original (Kimono)

(b) Proposed (Kimono)

(c) Original (DanceKiss)

(d) Proposed (DanceKiss)

Figure 8.30 Frame 77 SSIM heat map for low delay P encoded sequences Kimono and DanceKiss at 1Mbps



(a) Original (FlagShoot)

(b) Proposed (FlagShoot)

(c) Original (BQTerrace)

(d) Proposed (BQTerrace)

Figure 8.31 Frame 77 SSIM heat map for low delay P encoded sequences FlagShoot and BQTerrace at 1Mbps

8.7 Subjective testing results

The subjective testing results refer to where the proposed hybrid DCR-PC subjective test was applied. As described in Section 7.8, eight video sequences were shown (four random access, four low delay P), these sequences were run at four different bit rates of 16, 8, 4 and 2 Mbps. The test was taken by 11 participants, however, each run consisted of four random access and four low delay P encoded video sequences. This means through repeated measures baring one variable change, in this case bit-rate, the sample sizes per were 44 per participant, which in total is 352 observed findings. Which if divided by each configuration profile of random access and low delay P there are 176 observations each, then when divided by bit-rate within these profiles are 44 observations. This allowed the construction of Table 8.2 for each configuration and bit-rate, where they were analysed whether the proposed was subjectively inferior to the reference. This choice of analysis was because the previous PSNR results of pseudoSSIM in Section 4.4.2 were very poor during low delay P. In order to analyse whether subjectively the proposed PVC is inferior to the standard reference then a one-tail test is conducted, with a threshold (α) of 5%.

8.7.1 Understanding the subjective results

The results in Table 8.2 show statistical calculations based upon of their respective transformed responses. Each configuration of random access and low delay P is shown separately, divided by bit-rates and the respective statistical calculation. These statistical calculations provide a means to investigate that the proposed encoder is perceptually indistinguishable from the reference encoder. In particular, the use of confidence interval, p-value and power-test (also known as power of a test) are of significance to understand what they represent. Confidence interval, can indicate the accuracy of results recoded for the experiment. The p-value states likelihood or risk in predicting a sample based along a uniform distribution. It is the p-value in conjunction with α that a judgement of whether to accept or reject the null hypothesis. However, it is through the power-test that increased certainty can be placed in evaluating this p-value.

	Random access				
	16 Mbps	8 Mbps	4 Mbps	2 Mbps	Ave.
Sample mean	0.0682	0.0227	0.0682	0.0000	0.0398
Sample standard deviation	0.4523	0.4028	0.3339	0.3050	0.3735
Standard error of the mean	0.0682	0.0607	0.0503	0.0460	0.0563
Confidence interval (+/-)	0.1336	0.1190	0.0987	0.0901	0.1104
p-value	0.8385	0.6450	0.9086	0.5000	0.7230
Power-test: (1 - β)	0.8413	0.6458	0.9121	0.5000	0.7249

(a) Random access

	Low delay P				
	16 Mbps	8 Mbps	4 Mbps	2 Mbps	Ave.
Sample mean	0.0227	0.0000	0.0227	-0.0227	0.0057
Sample standard deviation	0.4028	0.3735	0.2631	0.2631	0.3257
Standard error of the mean	0.0607	0.0563	0.0397	0.0397	0.0491
Confidence interval (+/-)	0.1190	0.1104	0.0778	0.0778	0.0962
p-value	0.6450	0.5000	0.7152	0.2848	0.5362
Power-test: (1 - β)	0.6458	0.5000	0.7166	0.2833	0.5365

(b) Low delay P

Table 8.2 One tail t-test on subjective testing, the significant level was 5%, sample size was 44, and the null hypothesis was not rejected for both configurations and bit-rates.

The use of α value is to mitigate against a risk of misinterpreting an actual event where H_0 can not be rejected, as a statistical event where H_0 can be rejected. Where this occurs, it is known as a type I error and by increasing α, it reduces the risk of type I errors. The significant level, α is applied as the region of rejection and under a one tail test is more powerful than a two tail test as it concentrates the region of rejection on one side of a normalised uniform distribution. Conversely, there is the risk of H_0 is true when it should be false, meaning an event is considered non-rejection of the H_0 when it should be rejected. This is known as a type II error, β, the difference between sampled and actual population values. As the mean and population get closer β decreases, this is possible by increasing α and the sample size. Being able to obtain 1-β, produces the power-test which can provide the probability of correctly rejecting a false null hypothesis, where a value of above

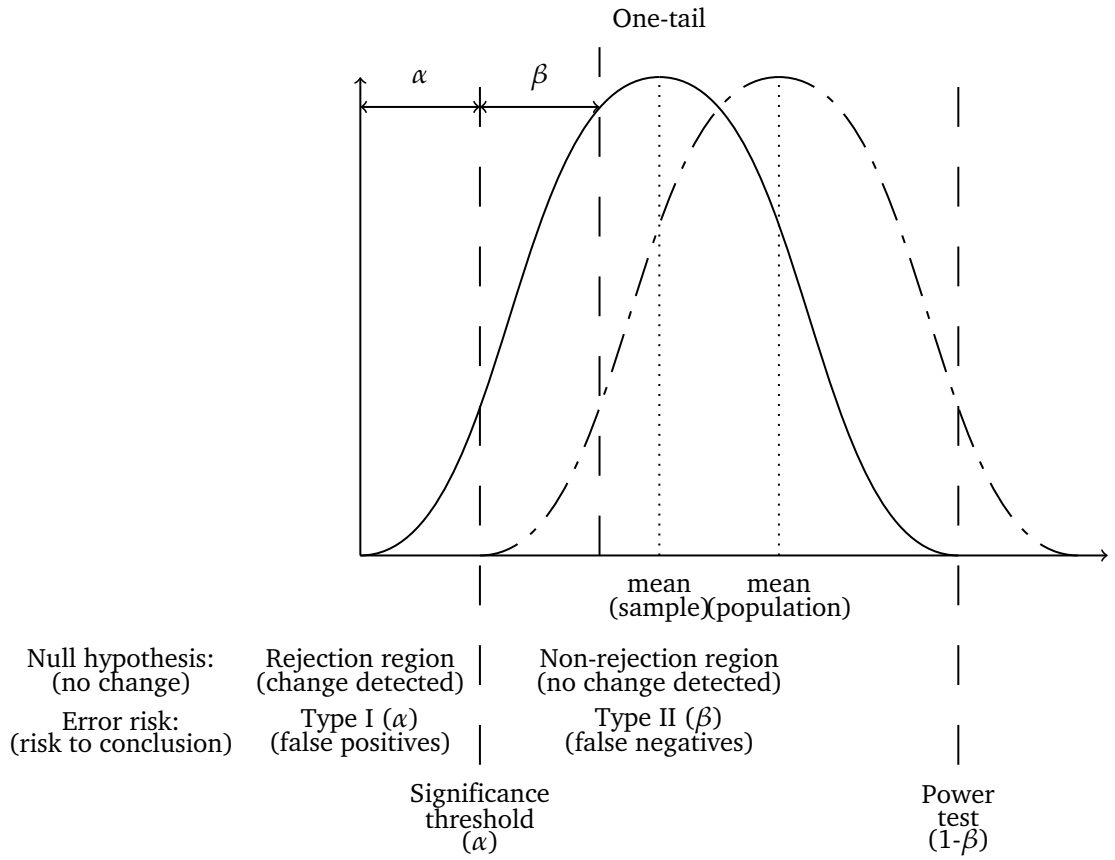


Figure 8.32 One-tail graph described with rejection and non-rejection regions with distribution of sampled (solid line) and population (dash-dotted)

0.8 represents where the sample and population means are close. This description of the one-tail analysis and risks in errors can be summarised in figure 8.32, where two distributions are shown, one representing the sampled and another reflecting the population .

Calculating the power-test depends upon having a standard deviation for the population. This can be estimated assuming normal distribution around a probable mean, in this case the population mean was considered to be zero.This value of zero was chosen to assume that if a control experiment had taken place of the same encoder being used in each video sequence shown, the results would say no difference throughout.

8.7.2 Analysing subjective results

The results in Table 8.2 were measured against a one-tail analysis with a significance of 5%, α. The proposed hybrid DCR-PC subjective test was designed

to provide five level responses of which encoding the participant preferred and by what degree. This meant that the one-tail test placed the 5% threshold on whether participants preferred the HEVC reference encoder or could not tell the difference between either encoder. Since the order of video sequences had been randomised, the results first needed to be transformed before conducting a one-tail test to produce a p-value. Where the p-value is less than α, then the null hypothesis (H_0) must be rejected, however, in the results p-value is $>> \alpha$, meaning that the H_0 can not be rejected. This was shown in both configurations across all bit-rates, meaning that the proposed were perceptually similar to the reference encoder for the same bit-rates. The power-test reinforced this scores similar to the respective p-values. Also, the confidence intervals were narrow illustrating that results were consistent.

As the p-value is greater than α, H_0 must be accepted, that subjectively there are no differences between the proposed or reference encoder. However, when examining these p-value results, under random access, the minimum is 0.5 and average is 0.72, while for low delay P it is 0.28 and 0.53 respectively. This suggests that along with earlier results, the changes in bit-redistribution, observed particularly in random access, is seen to be favourable. Perhaps following some further development to address the short comings in low delay P, it may be worth pursuing a subjective test experiment to identify whether the proposed encoder is subjectively favourable.

8.8 Summary of results

The results cover a range of tests, some gathered from existing log files, others where the decoder was modified to produce additional information and those from based upon participants. They were used for measuring objective image quality, complexity, bit allocation and subjective testing purposes. This meant a substantial set of results were presented in this chapter to illustrate the proposed hybrid STDM-IQA framework performance relative to the standard reference HEVC encoder. In the next chapter these results will be evaluated in terms of the design goals set and existing PVC solutions.

Chapter 9

Discussion of results

The discussion of results chapter will evaluate whether the design goals have been met and how this compares to other solutions. Overall, testing was driven to highlight the low complexity design, the use of perceptual assessment/activity to influence bit distribution and how this would affect subjective performance. In this chapter, these aspects will be covered based upon these results, from which an understanding of whether the proposed design and the implemented encoded operate as intended. This also means comparing the results against existing PVC solutions, however, this must be gauged in context of the research goal of producing a low complexity in-loop PVC solution.

9.1 Low complexity

The results in terms of overall timing show the complexity of the proposed encoder can be up to double figures in percentage terms, which is not ideal, yet almost tolerable. From the design stage, the proposed hybrid STDM-IQA framework extended the existing STDM only by offering a perceptual IQA path where conditions are met. Underlying this is the use of thresholds which govern when and were the IQA path should be undertaken. These threshold values

were initially set at the median of captured data that had been modelled, then later when simulating were refined to trigger on boundary, lighting changes or textures. The simulation results showed limited circumstances where the IQA path would be applied. As stated before, the sub-block level is highly intensive especially during prediction, in that context, the 6 to 11% additional overhead is satisfactory for the proposed encoder. This indicates that the perceptual pre/post check are being called to limit complexity, however, it does not indicate whether the perceptual IQA score has been added or not to the STDM. Regardless, the low complexity evaluation and content based perceptual triggering observed during the design stage of the hybrid STDM-IQA suggest that they are implemented with low complexity in mind as designed.

From the previous experiment of a sub-block PVC solution based on SSIM, the scaling of complexity was addressed to allow SSIM to occur at the sub-block level, however, the SSIM equation was still highly complex. During the SSIM based sub-block PVC solution, the overall complexity was typically near +20% and this operated only at the prediction stage. For the proposed hybrid STDM-IQA based PVC solution, the timing is far less and operates at all front-end stages, which demonstrates far lower complexity. There are other PVC solutions which offer lower overall complexity, yet these have a non-native design and are unable to offer individual candidate assessment (Yeo, H. L. Tan and Y. H. Tan, 2013; P. Zhao et al., 2013). The proposed encoder timing is highly competitive, half that of a JND based PVC solution, which also claims a low complexity design, however, it does have perceptual quantisation (J. Kim, Bae and M. Kim, 2015). In that solution the additional timing has an average of +22% and +11% for random access and low delay P respectively. This means that the hybrid STDM-IQA framework offers a PVC solution which has a lower computational burden than existing solutions.

9.2 STDM and perceptual differences

An important issue that was raised during the beginning of each contribution chapter was the differences between SSIM and STDM scores. This was shown at the sub-block level, with differences based either on selected samples or via a collection of observations. In either case their respective SSIM and STDM scores highlighted

that they would not correspond to a linear equation. This led to investigations which revealed that SSIM and STDM can be linked with covariance. In turn, this formed the basis for the SASD IQA for mode decision, which is part of the hybrid STDM-IQA PVC solution. Furthermore, being able to illustrate that covariance and JND had similar behaviour. This allowed combining SSIM luma and JND to produce APC and ppwAPC IQAs. These proposed IQAs provide a pixel-level means to evaluate distortion perceptually with a compatible score. However, these perceptual costs are applied given conditions are met, which are defined by the thresholds set during design and simulation stages and were designed to balance between complexity and perceptual significance.

From the results for SSIM, SSE and PSNR in Figures 8.1 and 8.2, the losses in STDM for the proposed are proportionally higher than that of 1-SSIM losses. An example is random access at 1Mbps where the loss is approximately -0.4 dB in PSNR and 0.005 in SSIM terms. This confirms that these proposed IQAs are able to minimise the losses under SSIM, while distortion is being introduced to the video sequence. This suggests that the use of these thresholds as part of the proposed PVC solution is providing a form of perceptual quality control. In the previous experiment of SSIM based sub-block prediction assessment, under low delay P profile, the losses for PSNR and SSIM were up to 1/3. The proposed hybrid STDM-IQA framework use of thresholds acts as a means regulate where perceptual cost can introduce non-HVS sensitive distortion. This is shown by the results where absolute distortion can be increased (which reduces PSNR), while perceptual losses are minimised (in SSIM).

These thresholds used to regulate when to apply perceptual cost are based upon captured and simulated data during the modelling and simulation stages. This involved three different videos, RaceHorses, CrowdRun and BasketballDrive, with the last two being of HD resolution. While this has been shown to be effective at choosing where to apply perceptual cost, it may not be the optimal threshold values given the limited video resources used. This means further modelling and simulation is required to determine whether a single, multiple or formula based thresholds need to be applied which could be activity and/or content based.

However, the choice of applying thresholds must balance against the potential of additional complexity overhead.

9.3 Bit redistribution by numbers

The proposed encoder is designed to encourage the redistribution of bits and this can be measured by the decoder analyser logs, which provides a total by block size width and those blocks with signalling only. This provides bit information for the entire video sequence, however, to extend this by frame the decoder was modified. From the above discussion in perceptual losses, the likelihood of changes was shown to be far higher for random access and particularly at lower bit-rates. The results in Figure 8.3 and Table 8.1 show that significant changes in block sizes are observed in random access. This is shown clearly with the scale of Δ bit usage for low delay P, $1/1000^{th}$ compared to random access in Figure 8.3. The figure illustrates that random access is able to reduce bits allocated to signalling with medium and larger block widths, while low delay P oscillates around zero. Existing SSIM based PVC solutions are able to extend perceptual bit-redistribution, to bit reduction, achieving bit savings and perceptual coding gain (Y.-H. Huang, Ou, Su et al., 2010; Yeo, H. L. Tan and Y. H. Tan, 2013). Similarly, a JND based PVC solution has demonstrated an average bit reduction of 16% for similar video quality on HEVC (J. Kim, Bae and M. Kim, 2015). While these PVC solutions offer bit reduction, they lack a native sub-block level design, to assess candidates individually, instead they are bounded by a block based model.

When examining by frame, Table 8.1 the changes within and between configurations are less pronounced as shown in Figure 8.3. In Table 8.1, low delay P at 1Mbps is similar to random access at 16Mbps, indicating that the changes in low delay P at 16Mbps are minimal. In random access 1Mbps, where these changes are significant, the differences suggest that signalling is increased by reducing bits for medium to large block widths of 16 and 32. While for random access 16Mbps and low delay P at 1Mbps, the changes occur at the same places of signalling, block widths of 16 and 32 with less magnitude, though not necessarily in the same direction. This suggests that video content rather than bit-rate can determine whether the IQA path is chosen. This indicates that the thresholds may be too

high for higher bit-rates in random access and low delay P across all bit-rates. A more dynamic threshold based upon configuration, bit budget or quantisation could enable greater perceptual bit-redistribution, encouraging medium, large or skip sub-blocks.

9.4 Bit redistribution via visual VCL tool

The visual VCL tool was designed to allow the visual analysis of the video coding layer (VCL) on the encoded bitstream, which superimposes the meta information on to the original video frame. This allows the distribution of meta information to be visually analysed on whether the bit-redistribution is occurring. In these results, residual QP and bit usage per LCU are shown in Figures 8.4 to 8.11 and Figures 8.12 to 8.19 respectively. The residual QP heat map shows on which regions the encoder considers important enough that it should store sub-block pixel residual information for a given bit-rate. Conversely, bit usage per LCU is where bits including signalling information is shown. This is a fairer representation of the spread of bits across the frame. Differences shown here from the reference encoder indicate that the proposed hybrid STDM-IQA framework is performing bit-redistribution in perceptually significant regions.

9.4.1 Quantised residue heat maps

For the proposed encoder, the residual QP heat maps illustrate that the hybrid STDM-IQA framework can influence sub-blocks choices in that they to wrap around the content of perceptual significant regions. This occurs both in random access and low delay P, yet it is more likely to occur in random access and during 1Mbps, such as in ParkScene and Tennis, Figures 8.4 and 8.5 respectively. In ParkScene, the tree and its branches produce a high contrast between the foreground and background, which most probably influenced the proposed encoder. For Tennis, the encoder attempts to wrap around the tennis players and the shadows on the tennis court floor as there is texture or boundary. When the perceptual significant and homogeneous regions are distinctive like in Figure 8.6, this affects bit-rates in different ways. At the lower bit-rate tighter wrapping is seen around the leg, while at the higher bit-rate, less spurious residual is allocated to homogeneous

regions. This is significant, as it shows that the proposed encoder can influence mode decision/RDO choices based upon content and bit-rate.

9.4.2 Bit usage per LCU

Another means to evaluate the perceptual bit-redistribution is to consider the bit usage per LCU, which includes bits allocated to for signalling. Under the high bit-rate of 16 Mbps the proposed encoder demonstrates its ability to redistribute bits by lowering bit usage for dark regions which are of medium or low textures. This is shown in ParkScene and Tennis in Figure 8.12, while for Pedestrian, it continues the suppression of spurious perceptually homogeneous regions as shown in Figure 8.13. For low delay P the changes are more subtle by the arrangement or expansion of regions. In Figure 8.14, the bit usage is high throughout except for the actor wearing the Kimono dress, where the LCUs are coloured in blue or green. The proposed compared to the reference have subtle differences, these include where dress folds and the ear of the actor. Compared to the reference encoder, for the homogeneous dark background in DanceKiss, the proposed encoder allocates more LCUs with zero or one bit. This demonstrates that the hybrid STDM-IQA framework encourages larger/skip blocks in dark areas that have a high JND threshold sensitivity. Likewise, when luma intensity and activity is high, when the JND threshold sensitivity is lower, which can mean a higher likelihood of the IQA path being used. This is shown in FlagShoot and BQTerrace in Figure 8.15, the proposed encoder tries to cover more with LCUs with the highest bit usage. For BQTerrace, this is a struggle for both encoders, with the proposed offering more LCUs with zero or one bit than the original reference encoder, thus allowing more bits to be applied elsewhere. This approach allows the proposed encoder to offer more coverage of the glass fence on the terrace where reflections and shadows are more likely.

For the 1Mbps LCU results, the proposed encoder lowers the LCU bit usage of perceptually homogeneous regions, which means more LCUs with zero or one bit allocated are shown as greyscale. The video sequences encoded under random access, Figures 8.16 and 8.17, the proposed encoder places a greater concentration of bits on fewer LCUs which are deemed perceptually significant. For low delay P the differences are again more subtle with changes occurring where

perceptual regions exist. In Figure 8.18, Kimono and DanceKiss the backgrounds are perceptually significant and homogeneous respectively, this leads to fewer zero or one bit LCUs in Kimono and more empty LCUs in DanceKiss. For active video sequences of FlagShoot and BQTerrace in Figure 8.19 like under 16Mbps in Figure 8.15, the proposed encoder tries to offer broader coverage of the perceptually significant regions of the various flags and across the entire terrace respectively.

Overall, the visual VCL tool demonstrate that the proposed encoder is able to allocate more bits towards perceptually significant areas and less for homogeneous regions. This ability of the proposed encoder to affect both the sub-block QP residue and LCU bit usage is observed for both configurations and bit-rates, however, this is greatest during random access at 1Mbps. For other bit rates and configurations the changes are more subtle, yet sufficient enough to be discussed. Comparing the results with SSIM based prediction sub-block experiment conducted earlier, similar behaviour is visually observed. In these limited results perceptually homogeneous regions have large sub-block, which suggested that fewer bits are allocated, which is in-line with PVC principles of perceptual redistribution. For the proposed encoder, the limited effect during low delay P needs to be resolved as this represents live video streaming applications. One approach is to lower the thresholds, this would potentially risk increasing the number of false positives and raise overall complexity. Therefore, extending the use of thresholds to have a greater impact on low delay P, is a balance of image quality and complexity. Also, bit distribution is an aspect of PVC solutions, however comparing with other existing solutions is difficult as information is not available. However, the VCL Tool used here can be used by with any HEVC compatible encoded bitstreams, which allow this to be possible if given access to the respective bitstreams.

9.5 Activity assessment simulation via visual VCL tool

Another aspect for discussion is the perceptual activity assessment, which as part of the hybrid STDM-IQA framework provides an additional perceptual score when the IQA path is chosen. When a perceptual cost is added, this indicates that a greater proportion of bits should be allocated to these regions, which should lead to lower quantisation in those sub-blocks. The results shown in Figures 8.20 to 8.27 compare

the original rate-control Hadamard 8x8 (RC Had 8x8) against the JND perceptual model and the proposed ppwAPC. From these results ppwAPC demonstrates it is able to trigger on those 8x8 pixel blocks which are bright and textured, suggesting they are of perceptual significance. This is shown with ParkScene and Tennis in Figures 8.20 and 8.21, where the trees (foreground) meet the sky (background), and the reflection on the umpires chair respectively. However, as a fixed 8x8 filter is applied it is liable to trigger when aliasing or a moire effect occurs as shown on the upper left quadrant in Figure 8.27, similarly on the actors in DanceKiss in Figure 8.25.

Compared to existing rate-control RC Had 8x8 and JND, ppwAPC is triggered on boundaries or shadows where there is a high contrast, which is of perceptual interest for the HVS. When these same regions are considered interesting by the alternatives of RC Had 8x8 and JND they cover a far wider area, than the trigger points of ppwAPC. The ppwAPC trigger point on the heat map tends to be red, however, the ppwAPC algorithm does provide lower scores on occasion, usually as orange or even green. In all, the colour range of ppwAPC is virtually binary, which compared to the tri-state of RC Had 8x8 and multicolour of JND, meaning that when ppwAPC is applied where perceptual activity is high, allowing the distinction to be more pronounced. The high cost applied by ppwAPC contributes towards the proposed encoder allocating smaller block sizes to the those regions which are can act as perceptual clues for the HVS. As rate-control is based upon the original frame and is a spatial only assessment, ppwAPC affects the proportion of bit-budget allocated to which parts of the frame and has no direct effect on bit-rate or quantisation. A more complex system which extends ppwAPC to involve quantisation could address this, allowing different ppwAPC responses depending upon bit-rates and coding types. In keeping with the region of interest (ROI) applications, those areas which are perceptually homogeneous could be quantised further to encourage bit savings or greater bit-redistribution. This would be suitable for video communication applications that have static or out of focus backgrounds, giving weighting to bright textured regions.

9.6 Distortion assessment at 1Mbps via visual VCL tool

The use of perceptual distortion assessment heat maps at 1Mbps in Figures 8.28 to 8.31 is to evaluate the effect of bit-redistribution via the proposed encoder. These results illustrate that the proposed encoder behaves similar to the original for both random access and low delay P. There is some distortion increase in particular background areas for ParkScene and Tennis as shown in Figure 8.28. For ParkScene there is high distortion present between the cyclist's left arm and their body, while for Tennis on the right hand side of the frame more distortion is present on the tennis court fence. A more subtle example is shown in PedestrianArea Figure 8.29, where greater distortion is added to the perceptually homogeneous wall on the right hand side of the frame, suggesting it would be tolerated by the HVS. For low delay P there are very minor differences, which reinforces the other results that low delay P has little perceptual bit-redistribution. This means that under random access the proposed encoder will retain perceptual clues while being more tolerant of distortion in perceptually homogeneous regions.

It should be noted that the SSIM equation is able to mask certain levels of pixel differences where the statistical calculations are identical (Fei et al., 2012). With this in mind, these results of the proposed encoder suggest that the distortion added is non-uniformly spread, particular in perceptually homogeneous regions where bit saving for similar distortion is possible. This would explain how in the previous experiment, of SSIM based sub-block prediction, there were substantial bit PSNR and SSIM losses under low delay P. These same reductions are not being repeated in this experiment, because the hybrid STDM-IQA framework regulate where a perceptual cost is added by performing perceptual significant tests. This suggests that the proposed encoder is balancing the perceptual picture quality throughout the frame. However, there is potential for future development, to investigate greater bit savings by applying perceptual quantisation.

9.7 Video frame texture

An issue that has been raised by perceptual video coding evaluation is the need for ground truths, however this is time consuming process (Chandler, 2013). While this has not been done here, the given original video sequence frame has been

analysed for its perceptual significant regions and discussed against the results in VCL changes observed. As this is done guided by edge detection than solely of a HVS, this process is first explained then the results are analysed.

The findings for bit-redistribution were analysed based upon a single frame for the respective video sequence. Within each of these frames, regions were highlighted to signify where the proposed encoder would operate differently to reference encoder. Ideally, a ground truth of the frame should be made which can be compared against, while this has not been done, their texture can be analysed by applying an edge detection filter. To assist the respective frames have been extracted from the uncompressed video sequence and processed using a filter in IrfanView called 'finding edges', set to a value of 3 (Skiljan, 2016). These processed frames are shown in Figures 9.1 and 9.2, and these figures will be examined in terms of the edge texture as a guide for perceptual significance.

9.7.1 Edge textures random access encoded video frame

In Figure 9.1 areas highlighted by ParkScene reflect high, medium and low areas of texture, from left to right. While the other respective frames of the video sequences, the highlighted regions are a mixture due to the camera movement or activity. For Tennis, the umpires chair is well defined, however, the movement by the actors/players has a mixed response. In comparison, in PedestrianArea, the high density of people moving means that they are only partially captured by the proposed encoder. Finally, the content of Riverbed demonstrates the strong edges incoherently, other than the ripple effect that is shown. When comparing these same highlighted regions with the sub-block QP partitioning in Figures 8.4 to 8.7, the proposed encoder indicates it is able to apply smaller block sizes on textured regions.

9.7.2 Edge textures low delay P encoded video frame

In Figure 9.2, the highlight regions can represent different types of textures. For Kimono, this is both low and no edge regions of the Kimono dress and the flowered background, while in DanceKiss, the feathered scarf is densely textured. For FlagShoot, the simple contrast in flag colours are distinctive and in BQTerrace, the metal railing is very clear, while the furniture and reflection is faint. Again,

as stated before comparing low delay P results are difficult since little change is apparent. This does indicate that these video sequences have similar levels of texture to those used in random access, yet the proposed encoder is unable to operate in a similar way to those video sequences encoded in random access.

(a) ParkScene

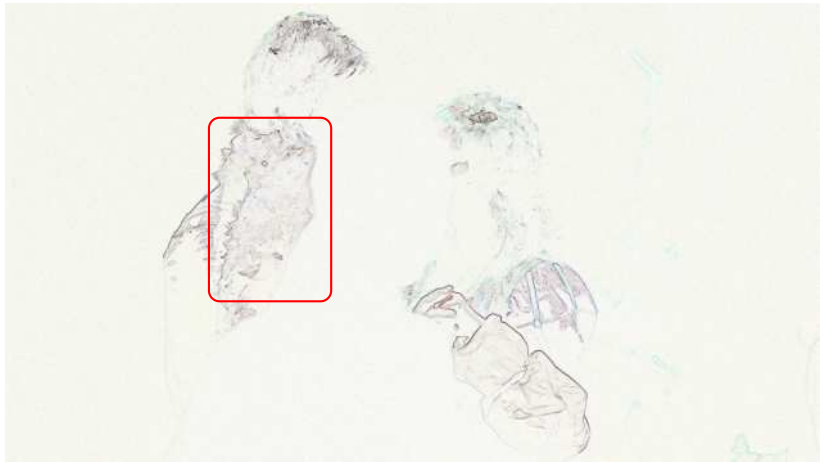
(b) Tennis

(c) PedestrianArea

(d) Riverbed

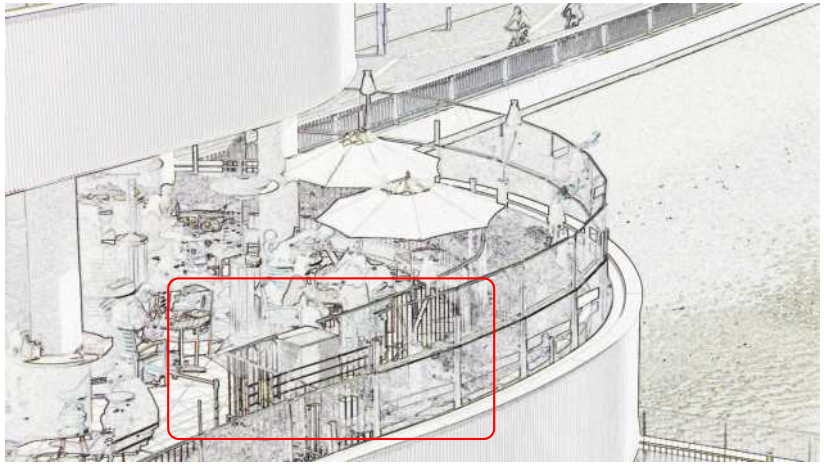
Figure 9.1 Frame 77 from uncompress video sequences with IrfanView filter ‘finding edges’ setting 3

(a) Kimono



(b) DanceKiss

(c) FlagShoot



(d) BQTerrace

Figure 9.2 Frame 77 from uncompress video sequences with IrfanView filter ‘finding edges’ setting 3

9.8 Subjective Testing

The use of subjective testing was to understand whether the changes detected for the proposed encoder during objective and visual tests would impact the HVS. Overall, the results in Table 8.2 shows that changes by the proposed encoder is subjectively similar to the original encoder. This suggests that these changes which included additional distortion, due to bit-redistribution and different partitioning was tolerated by the HVS. Across both configurations, the 95% confidence interval for the samples was approximately 0.1, which mean that for both random access and low delay P the sample mean could be negative or positive. This suggests that when the variability is accounted for in these results it is safer to conclude that there is no subjective difference detected. However, when comparing by configuration, the random access results suggest a stronger likelihood that it is no worse than the original, with p-values and power-test scores being significantly higher than low delay P. Also, while the sample mean tended around zero for both, it was likely to be higher for random access. This suggests that should the thresholds for low delay P be lowered, then more candidate choices could be influenced, which may result in similar findings to random access. In turn, this could extend random access profile to influence more regions, which could warrant investigating the subjective performance of whether the proposed is better than the reference.

This experiment builds upon the previous SSIM based sub-block prediction solution, which had substantial picture quality losses in low delay P. As this hybrid STDM-IQA experiment is more stable with far fewer SSIM losses, this suggest that the hybrid STDM-IQA approach is more favourable. Examining the random access results in Table 8.2a, there is an unusual behaviour for 16 and 4 Mbps verses 8 and 2 Mbps which leads to different range of p-values scores. The sample size was 44 per configuration per bit-rate, it was considered a sufficient figure at the time, yet undertaking a power analysis based upon the standard deviation may reveal otherwise. A future subjective study of this proposed encoder should conduct a power analysis, with an aim to apply a two-tail paired t-test attempt to prove a perceptual coding gain.

9.8.1 Subjective testing results as a series of repeated measures

The limited number of participants, does bring some doubt interpreting result. However, considering that each participant viewed the same eight video sequences four times, with only one variation change between them, bit-rate, they can be considered as repeated measures. This is where, across an experiment is considered over a time period and only a minimal of changes occur among the variables. In this case the time period is very short, approximately 20 minutes per participant. Also, this does depend on the participant being suitable and representative sample of the population. That is why before accepting participants, they were provided with details of the experiment, assessed on their suitability and had a test run before performing the experiment. This meant that the screening and training was sufficient to normalise variation. In addition, the experiment was especially designed to avoid preference or bias to the first or second video sequence by randomising their order. The results did not show outliers, meaning the data gathered was coherent. Had there been significant variation, then more participants would have been required to understand the results more accurately.

9.9 Understanding the limited redistribution under low delay P

Overall, the proposed encoder showed the framework to be more successful for random access than for low delay P. This maybe because the development was based upon random access encoded video and assumed the thresholds developed could be applied to both configurations. However, while random access used both B and P frames, it uses far fewer P frames. This means that developing also low delay P encoding may provide a more robust coverage for candidates under low delay P. This could be because of the coding types, random access uses B frames compared to P frame in low delay P and can be described where:

- B frames are not usually used for referencing by subsequent frames and have greater potential to be compressed at a high rate.
- P frames are referenced, meaning they have more information that should be preserved and thus have less potential for compression.

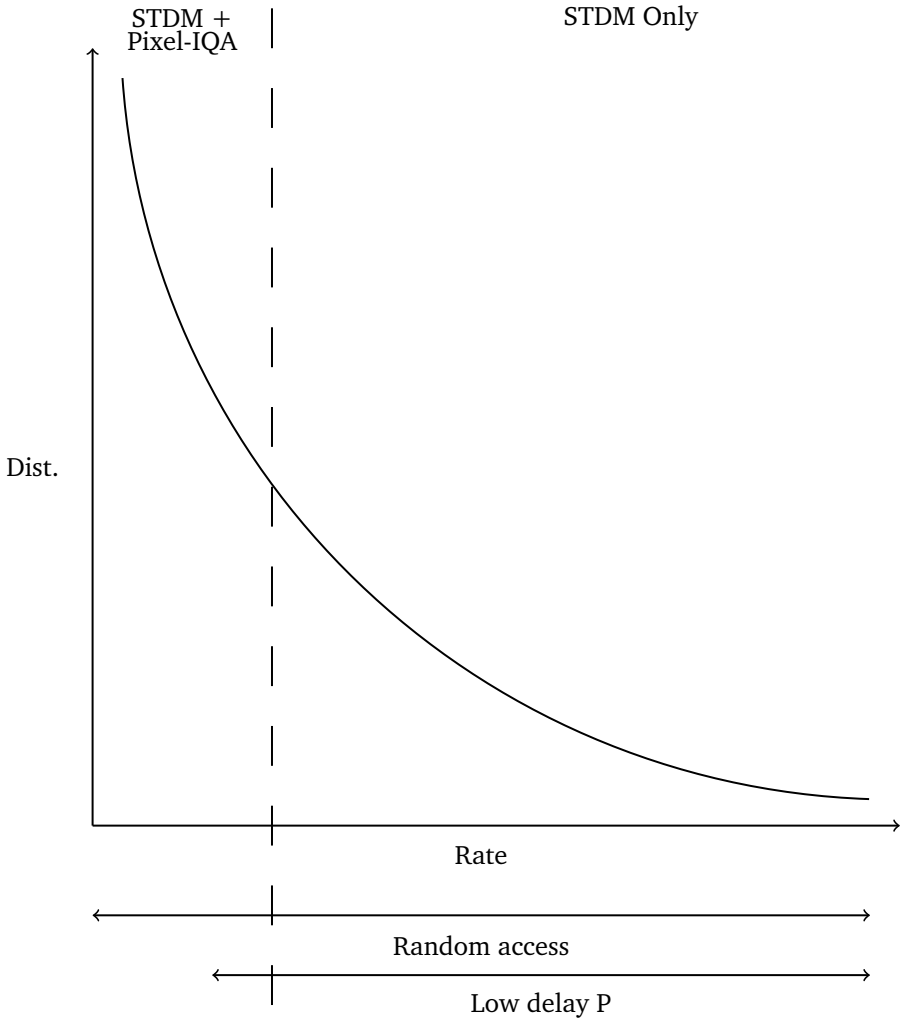


Figure 9.3 RD curve with pixel IQA by configuration

This issue can be illustrated via Figure 9.3, where current perceptual distortion threshold by which STDM undergoes with pixel-IQA is too high for low delay P to be applied. This means that currently, the changes detected are minor and most probably due to the rate-control activity which is quantisation independent and content based only. Lowering the thresholds used to encourage greater candidates to undergo the IQA path may risk complexity and score incompatibility. For very low distortion levels are virtually similar as shown in first series of investigation in this research. Therefore, a staggered approach is suggested between the high and low distortion extremes, as shown in Figure 9.3. This does mean covering a large proportion of the R-D curve space, with lower thresholds and smaller IQA costs. The pixel-IQA by way of the hybrid STDM-IQA framework have shown to be

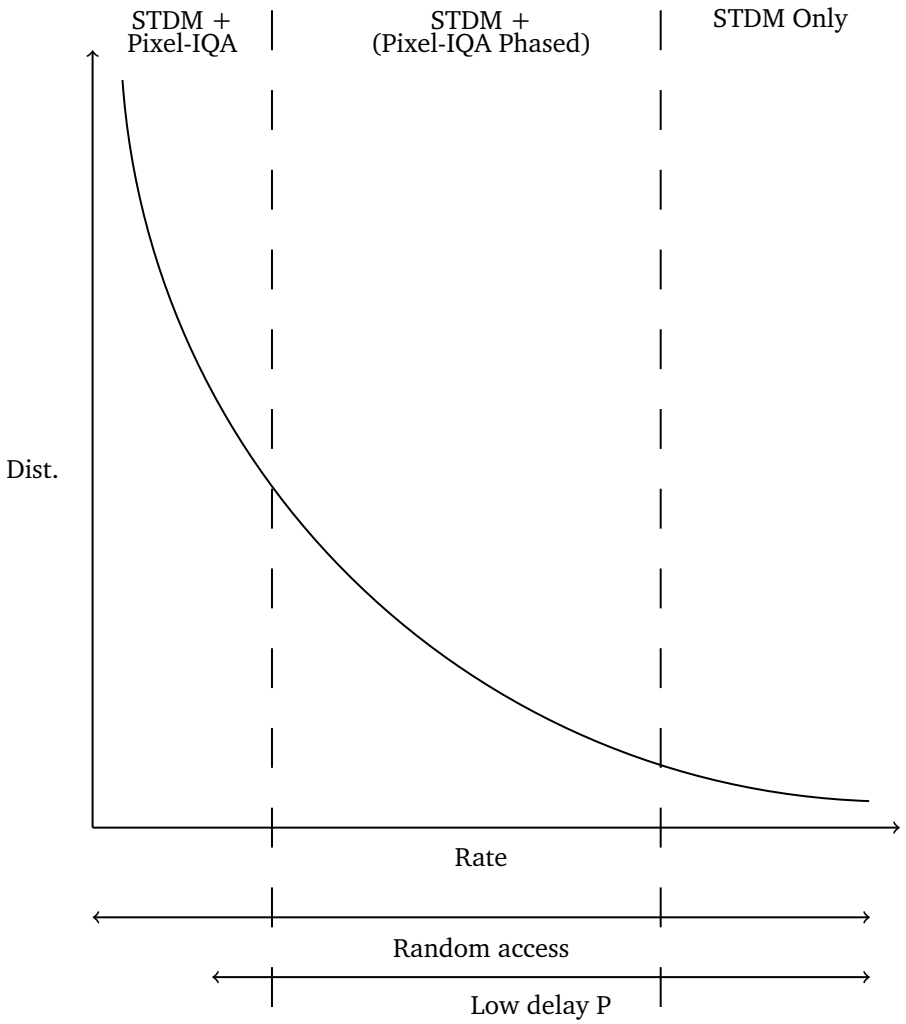


Figure 9.4 RD curve with pixel IQA by configuration with phased IQA score

effective at selecting perceptually significant regions. However, the development should be extended to gather low delay P encoded data for identifying whether the thresholds can be optimised. Another aspect is the need to add dimensionality of bit-rate in the development for both configuration profiles. Overall, in order to handle the variety of change, a more complex threshold based system requires to be developed.

9.10 Post-discussion findings

Following on from the above discussion, the proposed encoder was run on another, newer system based upon an Intel Core i5-6600K. This newer generation of processor was designed for HEVC, with full hardware support for motion

estimation and mode decision (Cutress, 2016). When Kimono was encoded under low delay P, both the difference in timing was reduced and bit-redistribution was shown to be more apparent. The proposed changes to the HEVC codebase were then ported to HM HEVC version 16.9 and the tests were re-run. However, as suggested, instead of encoding the entire video sequence, only half the video sequence was encoded (Moss et al., 2015). The findings illustrated that under low delay P bit-redistribution was occurring at similar levels to random access. This could be because the new processor has hardware support for HEVC encoding operations, allowing more candidates to be evaluated within a given period, which is significant for time critical profile of low delay P. These findings were then presented at the 2016 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) in Nara, Japan (Xplore, 2016). A copy of this paper is attached to the appendix A, which illustrates that given a processor optimised for HEVC encoding operations, the proposed changes are a low complexity means to redistribute sub-blocks.

The findings showed a timing overhead of $\approx 1/3$ compared to a PVC implementation based on JND (J. Kim, Bae and M. Kim, 2015). It led to a reduction in signalling bits of between 5 and 25 % and increases in bits for small, medium and large size sub-blocks. While for timing, this was $< +4\%$ and $< +6\%$ for low delay P and random access respectively. This timing is $\approx 1/3$ of existing PVC solution, which claims to be a low complexity PVC solution with perceptual quantisation (J. Kim, Bae and M. Kim, 2015). Overall, with optimised hardware, a low complexity in-loop PVC solution is possible, such that complexity it within single figures and bit-redistribution occurs to preserve perceptual integrity. However, further work needs to be done on investigating perceptual quantisation, with evaluation against a ground truth and subjective testing to identify any perceptual coding gain.

9.11 Summary of chapter

This chapter presented the discussion of results for the proposed low complexity sub-block level PVC solution in HEVC. The proposed hybrid STDM-IQA framework was designed to apply an IQA when there is significant perceptual distortion or activity. These test results indicated that changes, in terms of bit-redistribution were occurring on the video sequences encoded using random access than with

the low delay P. This was shown with lower bit usage in dark or low textures regions, which enabled more bits including partitioning for where perceptual clues for boundaries or high textures can be retained. Under subjective testing, bit-redistribution was accepted, however, under random access, the results indicated that further development may offer subjective preference. Importantly, when reflecting on the potential causes of the limited changes in low delay P, it was stated that threshold values may require revisiting. This is because the thresholds was developed based upon limited video sequences and using a single profile of random access at a fixed bit-rate. However, when the tests were run on a processor with HEVC hardware encoding support, it showed bit-redistribution is also occurring of equal measure on low delay P and random access. This suggests that time critical factor of low delay P may be restricting the volume of candidates assessed, which limits opportunities for the IQA path to be undertaken.

Chapter 10

Conclusion

Video compression is continuing being used in environments with limited computing and power resources in order to increase mobility and versatility of use. PVC needs to support these environments, especially as these applications are typical where potential gains will be beneficial for when video demand exceeds bandwidth availability. This issue is more likely as these applications involve low powered and portable devices, which are being produced at a rate of billions every year (Shimpi, 2014). Consequently, any PVC solution must be computationally friendly in order to overcome the issues of complexity typically associated with PVC (Chandler, 2013).

This research was an investigation to produce PVC using computationally friendly techniques. It meant first understanding existing STDM and SSIM relationship at the sub-block level before presenting a new approach to PVC, where alongside STDM a pixel-based IQA was called conditionally subject to meeting requirements. The first part highlighted that at the sub-block level covariance is a guide to associate SSIM with STDM scores. However, the complexity of SSIM and covariance led to a new approach of pixel based IQAs, as part of a proposed hybrid STDM-IQA framework. The design and development of this framework

involved using low complexity techniques and implemented in HEVC. The results highlighted that perceptual bit-redistribution is possible, and this was visualised using the proposed visual VCL tool. From these visualisations and log results bit-redistribution limited to those video sequences encoded using the random access profile. The subjective testing indicated that no perceived loss was experienced with bit-redistribution. However, when the proposed encoder ran on a processor optimised for HEVC encoding, bit-redistribution was seen in equal measure for both low delay P and random access profiles.

10.1 Research developments

From the existing literature, it was highlighted that there was limited understanding of perceptual assessment at the sub-block level. By exploring SSIM at this native sub-block level, the high complexity scaling typically applied for transforming SSIM to STDMs compatible scores was resolvable. This was done by illustrating that a common shared space, called a UBR, would be mapped based upon the SSIM component of covariance. The initial PVC solution called LHPSS was produced which demonstrated that a perceptual redistribution at the sub-block level. However, the complexity and instability of SSIM made it unsuitable, requiring a completely new approach, based upon low complexity perceptual assessment.

Since SSIM was designed for image coding, its emphasis was on perceptual ability before complexity. However, for video coding environment, IQAs need to be designed with low complexity in mind for them to be considered viable (Chandler, 2013). Existing SSIM based solutions are limited, and must manage the complexity SSIM introduces. Under SSIM, multiple mathematical operations are required to calculate the statistical properties of the image block, however, the SSIM score can averaged out or mask distortion (Fei et al., 2012). This means that in an SSIM based mode decision PVC solutions, given two statistically similar candidates, the one with the least bits it chosen, despite the very different STDM scores. This risks the loss of detail that may be perceptually significant (Zujovic, T. Pappas and Neuhoff, 2013). While for existing SSIM based PVC solutions overcome this with multiple overlapping window operations, this is not suitable for a low complexity solution. This led to the development of pixel-based IQAs, designed specifically for each

STDM norm space and added to the STDM cost if image loss by way of distortion is perceptually significant.

This new approach introduced the hybrid STDM-IQA framework, and the framework was modelled in R and then simulated visually on a video frame by producing the visual VCL tool. When implementing the framework on the HEVC encoder, the technical challenges meant that the original pixels were required and every sub-block combination need to be supported. The results indicated that bit-redistribution did occur while maintaining SSIM and a minor loss of PSNR. This was shown under those video sequences encoded under the random access profile. When visualising the VCL, the proposed encoder allocated smaller partitions and more bits to those areas which had high contrast, such as boundaries or bright textures. Equally, where these boundaries or textures were poorly defined, due to low contrast or low lighting, the proposed encoder would apply larger partitioning of sub-blocks and few bits. During subjective testing where a range of bit rates were shown, participants indicated that bit distribution was not considered detrimental.

Finally, the tests were re-run on a processor optimised for HEVC encoding and these showed substantial improvements, both in bit-redistribution and in timing. Under low delay P bit-redistribution happened at similar levels to random access, while preserving the integrity of perceptually significant regions.

10.2 Future work for this research

The proposed solution works well with an HEVC encoder optimised processor, however, the design was developed from data gathered under a single bit-rate (of 1Mbps), using limited video sequences and under random access only. To improve robustness of modelling, simulation and choice of thresholds, data should be collected across a broader set of bit rates, video sequences and also those encoded under low delay P. This will bring its own set of technical issues related to data management, due to the volume of prediction candidates produced. However, by extending the number of video sequences for developing and simulation, then the choice of thresholds can be evaluated under different conditions. Consequently, this may lead to a more complex threshold solution based upon content and encoding settings.

The proposed solution is able to offer in-loop sub-block level PVC, When comparing the proposed solutions with existing PVC solutions, they are able to offer perceptual quantisation (Dai et al., 2014; Y.-H. Huang, Ou, Su et al., 2010; Qi et al., 2013; Yeo, H. L. Tan and Y. H. Tan, 2013; Yuan et al., 2013). Perceptual quantisation can offer coding bit gain for similar perceptual quality, making the PVC solution attractive for limited bandwidth environments. For the proposed encoder to support perceptual quantisation would mean extending perceptual redistribution to increase quantisation on perceptually homogeneous regions and visa verse.

Finally, UHD is being adopted by broadcasters, content creators and manufactures, where increases in resolution, frame-rate, luma intensity and bit depth occur (Alliance, 2016). Where possible this should be supported as video sequence are being made available for testing labelled as HDR (University, 2016). This means extending the design beyond the 8 bit video and to be capable of handling 10 or even 12 bit video. Overall, this presents a technical challenge as pre-calulated values for the IQAs are stored in the LUTs. These LUTs will grow by a factor of four each time, for 10 and 12 bit video support, unless a memory efficient solution is found.

10.3 Closing thoughts

In video based communications, capturing and sharing is a popular activity, especially as the world online population continues to grow, and Internet access improves (Bank, 2016). Similarly, as heterogeneous networks evolve, such as 5G, the ability to support a variety of bandwidths, delays and distances increases (Dewar and Warren, 2014) This becomes more apparent with fixed or ad-hoc network infrastructure, such as machine to machine (M2M) and device to device (D2D) respectively, including those which combine fixed and ad-hoc, like vehicular to vehicular (V2V). This means that video based applications must adapt to changing conditions while supporting a variety of applications where the end user is the HVS. The existing approach to PVC is unattractive to low powered and/or portable devices, especially as the Internet of Things (IoT) is moving from hype to reality. PVC in these environments allow interaction and immersion to another space and/or time, however, as bandwidths and computing resources are put

under strain any solution must be aware of these constraints. Consequently, this research has rejected SSIM as part of native PVC solution and instead developed new pixel-based IQAs specific to respective norm space as part of a hybrid STDM-IQA framework. This new approach enables the potential for each candidate to be individually assessed with PVC at each front-end stage within a low complexity envelope which existing PVC solutions do not offer. The need for low complexity video coding is recognised as internet driven user generated content and broadcasts are forming projects to collaborate towards UHD (Codec, 2016; Cognitus, 2016). In addition, this research is designed for hybrid block-based encoders, allowing other video codecs to adopt the techniques discussed in this research and extend them further. This universal hybrid-block base design important as alternatives to HEVC and its successors are being developed as the world moves towards open real-time communications (Open Media, 2016; WebRTC, 2016).

References

[1] [Alliance, 2016]; U. Alliance. *UHD ALLIANCE PRESS RELEASEJanuary 4, 2016*. 2016. url: http://www.uhdalliance.org/uhd-alliance-press-releasejanuary-4-2016/#more-1227 (see pp. 243, 245).

[2] [AVSForum, 2016]; AVSForum. *AVS HD 709 - Blu-ray and MP4 Calibration - AVS Forum Home Theater Discussions And Reviews*. 2016. url: http://www.avsforum.com/forum/139-display-calibration/948496-avs-hd-709-blu-ray-mp4-calibration.html (see pp. 170, 245).

[3] [Avsforum, 2015]; Avsforum. *Comparing MPEG-2, H.264, and H.265 Video Codecs at NAB 2014 - AVS Forum | Home Theater Discussions And Reviews*. 2015. url: http://www.avsforum.com/forum/286-latest-industry-news/1528750-comparing-mpeg-2-h-264-h-265-video-codecs-nab-2014-a.html (visited on 14/11/2015) (see pp. 11, 245).

[4] [Bank, 2016]; W. Bank. *World Development Report 2016: Digital Dividends*. Tech. rep. World Bank, Jan. 2016. doi: 10.1596/978-1-4648-0671-1 (see pp. 243, 245).

[5] [Baruffa, 2016]; G. Baruffa. *DSPLab: PYUV: raw video sequence player*. 2016. url: http://dsplab.diei.unipg.it/software/pyuv_raw_video_sequence_player (see pp. 29, 245).

[6] [Bhat, Richardson and Kannangara, 2010]; A. Bhat, I. Richardson and S. Kannangara. "A New Perceptual Quality Metric for Compressed Video Based on Mean Squared Error". In: *Signal Processing: Image Communication* 25.8 (2010), pp. 588–596. doi: 10.1016/j.image.2010.07.002 (see pp. 36, 70, 71, 75, 79, 245).

[7] [Borer and Cotton, 2015]; T. Borer and A. Cotton. *A Display Independent High Dynamic Range Television System - BBC R&D*. 2015. url: http://www.bbc.co.uk/rd/publications/whitepaper309 (see pp. 2, 245).

[8] [Bossen et al., 2012]; F. Bossen et al. “HEVC Complexity and Implementation Analysis”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 22.12 (2012), pp. 1685–1696. doi: 10.1109/TCSVT.2012.2221255 (see pp. 3, 19, 245).

[9] [A. C. Bovik, 2013]; A. C. Bovik. “Automatic Prediction of Perceptual Image and Video Quality”. In: *Proceedings of the IEEE* 101.9 (2013), pp. 2008–2024. doi: 10.1109/JPROC.2013.2257632 (see pp. 34, 245).

[10] [Brooks, X. Zhao and T. Pappas, 2008]; A. Brooks, X. Zhao and T. Pappas. “Structural Similarity Quality Metrics in a Coding Context: Exploring the Space of Realistic Distortions”. In: *IEEE Transactions on Image Processing* 17.8 (Aug. 2008), pp. 1261–1273. doi: 10.1109/TIP.2008.926161 (see pp. 41, 58, 76, 96, 121, 124, 246).

[11] [Brunet et al., 2012]; D. Brunet et al. “Geodesics of the Structural Similarity index”. In: *Applied Mathematics Letters* 25.11 (2012), pp. 1921–1925. doi: 10.1016/j.aml.2012.03.001 (see pp. 56, 68, 70, 74, 246).

[12] [Bunkus.org, 2015]; Bunkus.org. *mkvmerge – Merge multimedia streams into a Matroska file*. 2015. url: https://www.bunkus.org/videotools/mkvtoolnix/doc/mkvmerge.html (visited on 13/11/2015) (see pp. 170, 246).

[13] [Caelli and Moraglia, 1986]; T. Caelli and G. Moraglia. “On the detection of signals embedded in natural scenes”. English. In: *Perception & Psychophysics* 39.2 (1986), pp. 87–95. doi: 10.3758/BF03211490 (see pp. 38, 246).

[14] [Carnec, Callet and Barba, 2008]; M. Carnec, P. L. Callet and D. Barba. “Objective Quality Assessment of Color Images based on a Generic Perceptual Reduced Reference”. In: *Signal Processing: Image Communication* 23.4 (2008), pp. 239–256. doi: 10.1016/j.image.2008.02.003 (see pp. 25, 33, 246).

[15] [Chandler, 2013]; D. M. Chandler. “Seven Challenges in Image Quality Assessment: Past, Present, and Future Research”. In: *ISRN Signal Processing* 2013 (Nov. 2013). doi: 10.1155/2013/905685 (see pp. 5, 25–27, 29, 38, 57–60, 97, 115, 229, 240, 241, 246).

[16] [M.-J. Chen and A. Bovik, 2010]; M.-J. Chen and A. Bovik. “Fast Structural Similarity index Algorithm”. In: *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. 2010, pp. 994–997. doi: 10.1109/ICASSP.2010.5495310 (see pp. 57, 246).

[17] [Chikkerur et al., 2011]; S. Chikkerur et al. “Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison”. In: *IEEE Transactions on Broadcasting* 57.2 (2011), pp. 165–182. doi: 10.1109/TBC.2011.2104671 (see pp. 35, 246).

[18] [Chou and Y.-C. Li, 1995]; C.-H. Chou and Y.-C. Li. “A Perceptually Tuned Subband Image Coder Based on the Measure of Just-Noticeable-Distortion Profile”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 5.6 (1995), pp. 467–476. doi: 10.1109/76.475889 (see pp. 31, 104, 246).

[19] [Cisco, 2015]; Cisco. *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014–2019*. Cisco. Feb. 2015. url: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html (see pp. 1, 246).

[20] [Codec, 2016]; T. Codec. *Turing codec, An open-source HEVC encoder*. 2016. url: http://turingcodec.org/ (see pp. 50, 244, 246).

[21] [Cognitus, 2016]; Cognitus. *Converging broadcast and user generated content for Interactive ultra-high definition services*. 2016. url: http://cognitus-h2020.eu/ (see pp. 50, 244, 246).

[22] [Cowdrick, 1917]; M. Cowdrick. "The Weber-Fechner Law and Sanford's Weight Experiment". English. In: *The American Journal of Psychology* 28.4 (1917), url: http://www.jstor.org/stable/1413900 (see pp. 25, 30, 246).

[23] [Cutress, 2016]; I. Cutress. *The Intel 6th Gen Skylake Review: Core i7-6700K and i5-6600K Tested*. 2016. url: http://www.anandtech.com/show/9483/intel-skylake-review-6700k-6600k-ddr4-ddr3-ipc-6th-generation/4 (see pp. 238, 247).

[24] [Dai et al., 2014]; W. Dai et al. "SSIM-based rate-distortion optimization in H.264". In: *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. 2014, pp. 7343–7347. doi: 10.1109/ICASSP.2014.6855026 (see pp. 41, 53, 57, 243, 247).

[25] [Demers, 2016]; C. Demers. *TV Size to Distance Calculator and Science*. 2016. url: http://uk.rtings.com/tv/reviews/by-size/size-to-distance-relationship (see pp. 54, 247).

[26] [Dewar and Warren, 2014]; C. Dewar and D. Warren. *Understanding 5G*. Tech. rep. GSMA Intelligence, Dec. 2014. url: https://gsmaintelligence.com/research/2014/12/understanding-5g/451/ (see pp. 243, 247).

[27] [D. W. Dong and Atick, 1995]; D. W. Dong and J. J. Atick. "Statistics of Natural Time-Varying Images". In: *Network: Computation in Neural Systems*. 1995, pp. 345–358. url: http://informahealthcare.com/doi/abs/10.1088/0954-898X_6_3_003 (see pp. 20, 247).

[28] [Y. Dong, M. T. Pourazad and Nasiopoulos, 2016]; Y. Dong, M. T. Pourazad and P. Nasiopoulos. "Human Visual System-Based Saliency Detection for High Dynamic Range Content". In: *IEEE Transactions on Multimedia* 18.4 (Apr. 2016), pp. 549–562. doi: 10.1109/TMM.2016.2522639 (see pp. 25, 247).

[29] [Duncan and Sarkar, 2012]; K. Duncan and S. Sarkar. "Saliency in Images and Video: A Brief Survey". In: *IET Computer Vision* 6.6 (2012), pp. 514–523. doi: 10.1049/iet-cvi.2012.0032 (see pp. 32, 34, 247).

[30] [Elder and Zucker, 1998]; J. H. Elder and S. W. Zucker. "Local scale control for edge detection and blur estimation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.7 (July 1998), pp. 699–716. doi: 10.1109/34.689301 (see pp. 49, 247).

[31] [Everett III, 1963]; H. Everett III. “Generalised Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources”. In: *Operations Research* 11.3 (1963), pp. 399–417. doi: 10.1287/opre.11.3.399 (see pp. 4, 16, 247).

[32] [Fei et al., 2012]; X. Fei et al. “Perceptual Image Quality Assessment based on Structural Similarity and Visual Masking”. In: *Signal Processing: Image Communication* 27.7 (2012), pp. 772–783. doi: 10.1016/j.image.2012.04.005 (see pp. 70, 96, 99, 101, 229, 241, 247).

[33] [Field, 1987]; D. J. Field. “Relations between the statistics of natural images and the response properties of cortical cells.” In: *Journal of the Optical Society of America. A, Optics and image science* 4.12 (Dec. 1987), pp. 2379–2394. doi: 10.1364/JOSAA.4.002379 (see pp. 38, 247).

[34] [Gabriellini, 2014]; A. Gabriellini. “HFR and video compression”. In: *EBU Tech-I Magazine* 19 (Mar. 2014), p. 7. url: https://tech.ebu.ch/files/live/sites/tech/files/shared/tech-i/ebu_tech-i_019.pdf (see pp. 2, 247).

[35] [Greenberg et al., 1997]; D. P. Greenberg et al. “A Framework for Realistic Image Synthesis”. In: *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH ’97. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1997, pp. 477–494. doi: 10.1145/258734.258914 (see pp. 59, 247).

[36] [Hong et al., 2010]; D. Hong et al. “H.264 Hierarchical P Coding in the Context of Ultra-Low Delay, Llow Complexity Applications”. In: *Proc. Picture Coding Symp. (PCS)*. 2010, pp. 146–149. doi: 10.1109/PCS.2010.5702445 (see pp. 169, 248).

[37] [Horé and Ziou, 2013]; A. Horé and D. Ziou. “Is there a Relationship between Peak-Signal-to-Noise Ratio and Structural Similarity index measure?” In: *IET Image Processing* 7.1 (2013), pp. 12–24. doi: 10.1049/iet-ipr.2012.0489 (see pp. 40, 63, 68, 74, 75, 96, 248).

[38] [Y.-H. Huang, Ou and H. Chen, 2010]; Y.-H. Huang, T.-S. Ou and H. Chen. “Perceptual-Based Coding Mode Decision”. In: *IEEE International Symposium on Circuits and Systems (ISCAS), Proceedings of 2010*. 302010-june2 2010, pp. 393–396. doi: 10.1109/ISCAS.2010.5537738 (see pp. 70, 248).

[39] [Y.-H. Huang, Ou, Su et al., 2010]; Y.-H. Huang, T.-S. Ou, P.-Y. Su et al. “Perceptual Rate-Distortion Optimization Using Structural Similarity Index as Quality Metric”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 20.11 (Nov. 2010), pp. 1614–1624. doi: 10.1109/TCSVT.2010.2087472 (see pp. 41, 49, 52, 53, 58, 79, 97, 224, 243, 248).

[40] [T. Huang, S. Dong and Tian, 2014]; T. Huang, S. Dong and Y. Tian. “Representing Visual Objects in HEVC Coding Loop”. In: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 4.1 (2014), pp. 5–16. doi: 10.1109/JETCAS.2014.2298274 (see pp. 97, 248).

[41] [ITU.int, 2015]; ITU.int. *BT.2020 Parameter values for ultra-high definition television systems for production and international programme exchange*. 2015. url: http://www.itu.int/rec/R-REC-BT.2020/en (see pp. 2, 248).

[42] [ITU-int, 2012]; ITU-int. *(BT.500) Methodology for the subjective assessment of the quality of television pictures*. 2012. url: http://www.itu.int/rec/R-REC-BT.500/en (see pp. 53, 170, 248).

[43] [ITU-T, 2008]; ITU-T. *P.910 (04/08) Subjective video quality assessment methods for multimedia applications*. ITU. Apr. 2008. url: https://www.itu.int/rec/T-REC-P.910-200804-I/en (see pp. 27, 28, 164, 165, 248).

[44] [JCT-VC HEVC, 2013]; B. JCT-VC HEVC. "Ticket No. 1212: SAD and HAD called during Inter where pixel values may be out of range, negative or exceed upper limit". In: *Hevc.hhi.fraunhofer.de* (2013). url: https://hevc.hhi.fraunhofer.de/trac/hevc/ticket/1212 (see pp. 154, 248).

[45] [JCT-VC, 2016]; F. H. H. I. JCT-VC. *High Efficiency Video Coding (HEVC) | JCT-VC*. 2016. url: https://hevc.hhi.fraunhofer.de/ (see pp. 154, 167, 248).

[46] [Jin and J. Chen, 2009]; R. Jin and J. Chen. "The Coding Rate Control of Consistent Perceptual Video Quality in H.264 ROI". In: *Proc. Int. Symp. Computer Network and Multimedia Technology CNMT 2009*. 2009, pp. 1–4. doi: 10.1109/CNMT.2009.5374597 (see pp. 43, 248).

[47] [Joshi, Loo, Shah, Rahman and Chang, 2013]; Y. Joshi, J. Loo, P. Shah, S. Rahman and Y. C. Chang. "A novel low complexity Local Hybrid Pseudo-SSIM-SATD distortion metric towards perceptual rate control". In: *Broadband Multimedia Systems and Broadcasting (BMSB), 2013 IEEE International Symposium on*. 2013, pp. 1–6. doi: 10.1109/BMSB.2013.6621695 (see pp. 7, 248).

[48] [Joshi, Loo, Shah, Rahman and Tasiran, 2015]; Y. Joshi, J. Loo, P. Shah, S. Rahman and A. Tasiran. "Low complexity sub-block perceptual distortion assessment for mode decision and rate-control". In: *Broadband Multimedia Systems and Broadcasting (BMSB), 2015 IEEE International Symposium on*. 2015, pp. 1–9. doi: 10.1109/BMSB.2015.7177262 (see pp. 7, 249).

[49] [Joshi, Shah et al., 2013]; Y. Joshi, P. Shah et al. "Review of Standard Traditional Distortion Metrics and a need for Perceptual Distortion Metric at a (Sub) Macroblock Level". In: *Broadband Multimedia Systems and Broadcasting (BMSB), 2013 IEEE International Symposium on*. 2013, pp. 1–6. doi: 10.1109/BMSB.2013.6621755 (see pp. 7, 249).

[50] [Julesz, 1981]; B. Julesz. "Textons, the elements of texture perception, and their interactions". In: *Nature* 290.5802 (Mar. 1981), pp. 91–97. doi: 10.1038/290091a0 (see pp. 36, 38, 249).

[51] [Kay and Lemay, 1986]; S. Kay and G. Lemay. "Edge detection using the linear model". In: *and Signal Processing IEEE Transactions on Acoustics, Speech* 34.5 (Oct. 1986), pp. 1221–1227. doi: 10.1109/TASSP.1986.1164941 (see pp. 31, 42, 249).

[52] [Kelly, 1979]; D. H. Kelly. "Motion and vision. II. Stabilized spatio-temporal threshold surface". In: *J. Opt. Soc. Am.* 69.10 (Oct. 1979), pp. 1340–1349. doi: 10.1364/JOSA.69.001340 (see pp. 30, 249).

[53] [Kilkenny et al., 2010]; C. Kilkenny et al. “Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research”. In: *PLoS Biol* 8.6 (June 2010), e1000412. doi: 10.1371/journal.pbio.1000412 (see pp. 170, 249).

[54] [J. Kim, Bae and M. Kim, 2015]; J. Kim, S.-H. Bae and M. Kim. “An HEVC-Compliant Perceptual Video Coding Scheme Based on JND Models for Variable Block-Sized Transform Kernels”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 25.11 (2015), pp. 1786–1800. doi: 10.1109/TCSVT.2015.2389491 (see pp. 49, 54, 222, 224, 238, 249).

[55] [Y. Kim et al., 2012]; Y. Kim et al. “A Rate-Distortion cost estimation approach to fast intra prediction in Video Coding for Ultra High Definition TV applications”. In: *Proc. IEEE Int Consumer Electronics (ICCE) Conf*. 2012, pp. 164–165. doi: 10.1109/ICCE.2012.6161790 (see pp. 49, 249).

[56] [Kolb, 2003]; H. Kolb. “How the Retina Works”. In: *Amer. Scientist* 91.1 (2003), p. 28. doi: 10.1511/2003.1.28 (see pp. 24, 249).

[57] [Le Callet and Niebur, 2013]; P. Le Callet and E. Niebur. “Visual Attention and Applications in Multimedia Technologies”. In: *Proceedings of the IEEE* 101.9 (2013), pp. 2058–2067. doi: 10.1109/JPROC.2013.2265801 (see pp. 32, 33, 35, 249).

[58] [Lee and Ebrahimi, 2012]; J.-S. Lee and T. Ebrahimi. “Perceptual Video Compression: A Survey”. In: *IEEE Journal of Selected Topics in Signal Processing* 6.6 (2012), pp. 684–697. doi: 10.1109/JSTSP.2012.2215006 (see pp. 2, 249).

[59] [B. Li, H. Li et al., 2012]; B. Li, H. Li et al. *Rate control by R-lambda model for HEVC (JCTVC-K0103)*. Tech. rep. University of Science and Technology of China, 2012. url: http://phenix.int-evry.fr/jct/doc_end_user/documents/11_Shanghai/wg11/JCTVC-K0103-v2.zip (see pp. 17, 249).

[60] [B. Li, D. Zhang et al., 2012]; B. Li, D. Zhang et al. *QP determination by lambda value (JCTVC-I0426)*. Tech. rep. University of Science and Technology of China Microsoft Corp., 2012. url: http://phenix.int-evry.fr/jct/doc_end_user/documents/9_Geneva/wg11/JCTVC-I0426-v3.zip (see pp. 17, 250).

[61] [J.-L. Lin et al., 2013]; J.-L. Lin et al. “Motion Vector Coding in the HEVC Standard”. In: *IEEE Journal of Selected Topics in Signal Processing* 7.6 (2013), pp. 957–968. doi: 10.1109/JSTSP.2013.2271975 (see pp. 20, 250).

[62] [W. Lin and Kuo, 2011]; W. Lin and C.-C. J. Kuo. “Perceptual Visual Quality Metrics: A Survey”. In: *Journal of Visual Communication and Image Representation* 22.4 (2011), pp. 297–312. doi: 10.1016/j.jvcir.2011.01.005 (see pp. 35, 39, 250).

[63] [Ma et al., 2011]; Z. Ma et al. “Modeling of Rate and Perceptual Quality of Video and its Application to Frame Rate Adaptive Rate Control”. In: *Proc. 18th IEEE Int Image Processing (ICIP) Conf*. 2011, pp. 3321–3324. doi: 10.1109/ICIP.2011.6116382 (see pp. 43, 250).

[64] [Mannos and Sakrison, 1974]; J. Mannos and D. Sakrison. “The effects of a visual fidelity criterion of the encoding of images”. In: *Information Theory, IEEE Transactions on* 20.4 (July 1974), pp. 525–536. doi: 10.1109/TIT.1974.1055250 (see pp. 25, 30, 250).

[65] [Matroska.org, 2015]; Matroska.org. *Matroska Media Container - Homepage | Matroska*. 2015. url: http://www.matroska.org/ (visited on 13/11/2015) (see pp. 170, 250).

[66] [Misra et al., 2013]; K. Misra et al. “An Overview of Tiles in HEVC”. In: *IEEE Journal of Selected Topics in Signal Processing* 7.6 (2013), pp. 969–977. doi: 10.1109/JSTSP.2013.2271451 (see pp. 57, 250).

[67] [Moorthy and A. Bovik, 2009]; A. Moorthy and A. Bovik. “A Motion Compensated approach to Video Quality Assessment”. In: *Signals, Systems and Computers, 2009 Conference Record of the Forty-Third Asilomar Conference on*. 2009, pp. 872–875. doi: 10.1109/ACSSC.2009.5469994 (see pp. 57, 250).

[68] [Moss et al., 2015]; F. M. Moss et al. “On the Optimal Presentation Duration for Subjective Video Quality Assessment”. In: *IEEE Transactions on Circuits and Systems for Video Technology* PP.99 (2015), p. 1. doi: 10.1109/TCSVT.2015.2461971 (see pp. 28, 238, 250).

[69] [Ndjiki-Nya et al., 2012]; P. Ndjiki-Nya et al. “Perception-oriented Video Coding based on Image Analysis and Completion: A Review”. In: *Signal Processing: Image Communication* 27.6 (2012), pp. 579–594. doi: 10.1016/j.image.2012.01.003 (see pp. 32, 33, 250).

[70] [Noland, 2014]; K. Noland. *The Application of Sampling Theory to Television Frame Rate Requirements - BBC R&D*. 2014. url: http://www.bbc.co.uk/rd/publications/whitepaper282 (see pp. 2, 250).

[71] [Open Media, 2016]; A. for Open Media. *Home*. 2016. url: http://aomedia.org/ (see pp. 5, 244, 250).

[72] [Ortega and Ramchandran, 1998]; A. Ortega and K. Ramchandran. “Rate-distortion methods for image and video compression”. In: *IEEE Signal Processing Magazine* 15.6 (1998), pp. 23–50. doi: 10.1109/79.733495 (see pp. 16, 250).

[73] [T. N. Pappas et al., 2013]; T. N. Pappas et al. “Image Analysis: Focus on Texture Similarity”. In: *Proceedings of the IEEE* 101.9 (Aug. 2013), pp. 2044–2057. doi: 10.1109/JPROC.2013.2262912 (see pp. 10, 35, 38, 251).

[74] [Péchard et al., 2007]; S. Péchard et al. “A New Methodology to Estimate the Impact of H.264 Artefacts on Subjective Video Quality”. English. In: *Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics, VPQM2007*. Scottsdale United States, Jan. 2007, p. 373. url: http://hal.archives-ouvertes.fr/hal-00275334/en/ (see pp. 33, 251).

[75] [Peli, 2001]; E. Peli. “Contrast sensitivity function and image discrimination”. In: *J. Opt. Soc. Am. A* 18.2 (Feb. 2001), pp. 283–293. doi: 10.1364/JOSAA.18.000283 (see pp. 26, 115, 251).

[76] [E. G. Pereira and R. Pereira, 2015]; E. G. Pereira and R. Pereira. “Video Encoding and Streaming Mechanisms in IoT Low Power Networks”. In: *Proc. 3rd Int Future Internet of Things and Cloud (FiCloud) Conf*. Aug. 2015, pp. 357–362. doi: 10.1109/FiCloud.2015.88 (see pp. 3, 251).

[77] [Pinson, Janowski and Papir, 2015]; M. Pinson, L. Janowski and Z. Papir. “Video Quality Assessment: Subjective testing of entertainment scenes”. In: 32.1 (2015), pp. 101–114. doi: 10.1109/MSP.2013.2292535 (see pp. 27, 28, 165, 251).

[78] [M. Pourazad et al., 2012]; M. Pourazad et al. “HEVC: The New Gold Standard for Video Compression: How Does HEVC Compare with H.264/AVC?” In: *Consumer Electronics Magazine, IEEE* 1.3 (July 2012), pp. 36–46. doi: 10.1109/MCE.2012.2192754 (see pp. 19, 251).

[79] [Project, 2015]; Q. Project. *Qt Creator - The IDE*. 2015. url: http://www.qt.io/ide/ (visited on 18/12/2015) (see pp. 168, 251).

[80] [Provision-itn.eu, 2015]; Provision-itn.eu. *Provision - Initial Training Network*. 2015. url: http://www.provision-itn.eu/home.htm (see pp. 2, 50, 251).

[81] [Qi et al., 2013]; J. Qi et al. “Efficient rate-distortion optimization for HEVC using SSIM and motion homogeneity”. In: *Picture Coding Symposium (PCS), 2013*. 2013, pp. 217–220. doi: 10.1109/PCS.2013.6737722 (see pp. 53, 243, 251).

[82] [R Core Team, 2014]; R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2014. url: http://www.R-project.org/ (see pp. 85, 133, 251).

[83] [R&D, 2012]; B. R&D. *IP Studio - BBC R&D*. 2012. url: http://www.bbc.co.uk/rd/projects/ip-studio (see pp. 2, 251).

[84] [Reinhard et al., 2013]; E. Reinhard et al. “On Visual Realism of Synthesized Imagery”. In: *Proceedings of the IEEE* 101.9 (2013), pp. 1998–2007. doi: 10.1109/JPROC.2013.2260711 (see pp. 59, 251).

[85] [Richter, 2013]; T. Richter. “A Global Image Fidelity Metric: Visual Distance and its Properties”. In: *Image Processing (ICIP), 2013 20th IEEE International Conference on*. 2013, pp. 369–373. doi: 10.1109/ICIP.2013.6738076 (see pp. 56, 251).

[86] [2011]; T. Richter. “SSIM as Global Quality Metric: A Differential Geometry View”. In: *Proc. Third Int Quality of Multimedia Experience (QoMEX) Workshop*. 2011, pp. 189–194. doi: 10.1109/QoMEX.2011.6065701 (see pp. 37, 56, 251).

[87] [Robin E. N. Horne, 1998]; S. J. S. Robin E. N. Horne, ed. *The Colour Image Processing Handbook*. Springer Berlin / Heidelberg, 1998. doi: 10.1007/978-1-4615-5779-1 (see pp. 31, 42, 252).

[88] [Rouse and Hemami, 2008]; D. Rouse and S. Hemami. “Understanding and Simplifying the Structural Similarity Metric”. In: *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. 2008, pp. 1188–1191. doi: 10.1109/ICIP.2008.4711973 (see pp. 57, 78, 98, 252).

[89] [Sanders, 2015]; J. Sanders. *Veusz - A Scientific Plotting Package*. 2015. url: http://home.gna.org/veusz/ (see pp. 85, 106, 252).

[90] [Schmolesky, 2016]; M. Schmolesky. *The Primary Visual Cortex by Matthew Schmolesky – Webvision*. 2016. url: http://webvision.med.utah.edu/book/part-ix-psychophysics-of-vision/the-primary-visual-cortex/ (see pp. 24, 25, 252).

[91] [Schödl et al., 2000]; A. Schödl et al. "Video Textures". In: *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '00. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 2000, pp. 489–498. doi: 10.1145/344779.345012 (see pp. 59, 252).

[92] [Shimpi, 2014]; A. Shimpi. *ARM Partners Ship 50 Billion Chips Since 1991 - Where Did They Go?* Mar. 2014. url: http://www.anandtech.com/show/7909/arm-partners-ship-50-billion-chips-since-1991-where-did-they-go (see pp. 240, 252).

[93] [Skiljan, 2016]; I. Skiljan. *IrfanView - Official Homepage - one of the most popular viewers worldwide*. 2016. url: http://www.irfanview.com/ (see pp. 230, 252).

[94] [Smith, 2016]; R. Smith. *ARM Announces Mali Egil Video Processor: VP9 Encode & Decode For Mobile*. 2016. url: http://www.anandtech.com/show/10428/arm-announces-mali-egil-video-processor (see pp. 3, 6, 252).

[95] [Su et al., 2012]; P.-Y. Su et al. "Adopting Perceptual Quality Metrics in Video Encoders: Progress and Critiques". In: *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*. 2012, pp. 73–78. doi: 10.1109/ICMEW.2012.20 (see pp. 53, 59, 60, 66, 97, 98, 100, 252).

[96] [Sühring, n.d.]; K. Sühring. *H.264/AVC Reference Software JM*. url: http://iphome.hhi.de/suehring/tml/ (see pp. 79, 83, 84, 133, 252).

[97] [Sullivan et al., 2012]; G. J. Sullivan et al. "Overview of the High Efficiency Video Coding (HEVC) Standard". In: *IEEE Transactions on Circuits and Systems for Video Technology* 22.12 (Dec. 2012), pp. 1649–1668. doi: 10.1109/TCSVT.2012.2221191 (see pp. 19, 23, 48, 86, 154, 252).

[98] [Thira, 2015]; Thira. *THIRA*. 2015. url: http://thira.ch.bbc.co.uk/ (see pp. 3, 50, 252).

[99] [University, 2016]; S. J. T. University. *SJTU 4K Video Sequences*. 2016. url: http://medialab.sjtu.edu.cn/web4k/index.html (see pp. 26, 29, 243, 252).

[100] [Videolan.org, 2015]; Videolan.org. *VideoLAN - VLC: Official site - Free multimedia solutions for all OS!* 2015. url: http://www.videolan.org/ (visited on 13/11/2015) (see pp. 170, 252).

[101] [VQEG, 2004]; V. Q. E. G. VQEG. *ITU-T Tutorial: Objective perceptual assessment of video quality: Full reference television*. ITU-T. 2004. url: http://www.itu.int/ITU-T/studygroups/com09/docs/tutorial_opavc.pdf (see pp. 27, 252).

[102] [H. Wang, Qian and Liu, 2010]; H. Wang, X. Qian and G. Liu. “Inter Mode Decision Based on Just Noticeable Difference Profile”. In: *Proc. 17th IEEE Int Image Processing (ICIP) Conf*. 2010, pp. 297–300. doi: 10.1109/ICIP.2010.5653462 (see pp. 44, 253).

[103] [Z. Wang et al., 2004]; Z. Wang et al. “Image Quality Assessment: From Error Visibility to Structural Similarity”. In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612. doi: 10.1109/TIP.2003.819861 (see pp. 10, 26, 34, 37, 39–41, 58, 104, 253).

[104] [Weber, 1864]; E. H. Weber. *Der Tastsinn und das Gemeingefühl*. 1864. url: http://www.uni-leipzig.de/~psycho/wundt/opera/ehweber/tastsinn/Tinhalt.htm (see pp. 25, 30, 69, 253).

[105] [WebRTC, 2016]; WebRTC. *WebRTC Home | WebRTC*. 2016. url: https://webrtc.org/ (see pp. 5, 244, 253).

[106] [Wiegand et al., 2010]; T. Wiegand et al. “Circuits and Systems for Video Technology, IEEE Transactions on”. In: *Issue: 12* 20.12 (2010), pp. 1661–1666. doi: 10.1109/TCSVT.2010.2095692 (see pp. 23, 253).

[107] [G.-L. Wu et al., 2013]; G.-L. Wu et al. “Perceptual Quality-Regulable Video Coding System With Region-Based Rate Control Scheme”. In: *IEEE Transactions on Image Processing* 22.6 (2013), pp. 2247–2258. doi: 10.1109/TIP.2013.2247409 (see pp. 75, 96, 253).

[108] [H. R. Wu, W. Lin and Ngan, 2014]; H. R. Wu, W. Lin and K. N. Ngan. “Rate-perceptual-distortion optimization (RpDO) based picture coding — Issues and challenges”. In: *Digital Signal Processing (DSP), 2014 19th International Conference on*. 2014, pp. 777–782. doi: 10.1109/ICDSP.2014.6900770 (see pp. 27, 28, 31, 253).

[109] [H. R. Wu, Reibman et al., 2013]; H. R. Wu, A. Reibman et al. “Perceptual Visual Signal Compression and Transmission”. In: *Proceedings of the IEEE* 101.9 (2013), pp. 2025–2043. doi: 10.1109/JPROC.2013.2262911 (see pp. 10, 30, 31, 35, 253).

[110] [H. Wu and Rao, 2005]; H. Wu and K. Rao, eds. *Digital Video Image Quality and Perceptual Coding*. CRC Press, 2005. doi: 10.1201/9781420027822.fmatt (see pp. 69, 79, 253).

[111] [T.-H. Wu, G.-L. Wu and Chien, 2009]; T.-H. Wu, G.-L. Wu and S.-Y. Chien. “Bio-inspired Perceptual Video Encoding based on H.264/AVC”. In: *Proc. IEEE Int. Symp. Circuits and Systems ISCAS 2009*. 2009, pp. 2826–2829. doi: 10.1109/ISCAS.2009.5118390 (see pp. 44, 104, 253).

[112] [Xiph.org, 2016]; Xiph.org. *Xiph.org Video Test Media [derf’s collection]*. 2016. url: http://media.xiph.org/video/derf/ (see pp. 29, 253).

[113] [Xplore, 2016]; I. Xplore. *IEEE Xplore - Conference Home Page*. 2016. url: http://ieeexplore.ieee.org/servlet/opac?punumber=1001845 (see pp. 7, 238, 253).

[114] [Yang et al., 2005]; X. Yang et al. “Motion-compensated residue preprocessing in video coding based on just-noticeable-distortion profile”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 15.6 (June 2005), pp. 742–752. doi: 10.1109/TCSVT.2005.848313 (see pp. 49, 253).

[115] [Yeo, H. L. Tan and Y. H. Tan, 2013]; C. Yeo, H. L. Tan and Y. H. Tan. “On Rate Distortion Optimization Using SSIM”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 23.7 (2013), pp. 1170–1181. doi: 10.1109/TCSVT.2013.2240918 (see pp. 49, 53, 57, 58, 96, 222, 224, 243, 253).

[116] [Yogeshwar and Mammone, 1990]; J. Yogeshwar and R. J. Mammone. “A New Perceptual Model for Video Sequence Encoding”. In: *Proc. Conf. th Int Pattern Recognition*. 1990, pp. 188–193. doi: 10.1109/ICPR.1990.119352 (see pp. 30, 31, 254).

[117] [Yu et al., 2005]; H. Yu et al. “A Perceptual Bit Allocation Scheme for H.264”. In: *Proc. IEEE Int. Conf. Multimedia and Expo ICME 2005*. 2005. doi: 10.1109/ICME.2005.1521423 (see pp. 42, 254).

[118] [Yuan et al., 2013]; Z. Yuan et al. “A perceptual rate-distortion optimization approach based on piecewise linear approximation for video coding”. In: *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*. 2013, pp. 1–5. doi: 10.1109/ICMEW.2013.6618280 (see pp. 53, 243, 254).

[119] [F. Zhang and Bull, 2015]; F. Zhang and D. Bull. “A Perception-based Hybrid Model for Video Quality Assessment”. In: *IEEE Transactions on Circuits and Systems for Video Technology* PP.99 (2015), p. 1. doi: 10.1109/TCSVT.2015.2428551 (see pp. 53, 59, 100, 115, 254).

[120] [P. Zhao et al., 2013]; P. Zhao et al. “Low-complexity content-adaptive Lagrange multiplier decision for SSIM-based RD-optimized video coding”. In: *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*. May 2013, pp. 485–488. doi: 10.1109/ISCAS.2013.6571886 (see pp. 222, 254).

[121] [Zujovic, T. Pappas and Neuhoff, 2013]; J. Zujovic, T. Pappas and D. Neuhoff. “Structural Texture Similarity Metrics for Image Analysis and Retrieval”. In: *Image Processing, IEEE Transactions on* 22.7 (July 2013), pp. 2545–2558. doi: 10.1109/TIP.2013.2251645 (see pp. 58, 101, 241, 254).

Appendix A

Published research

Review of Standard Traditional Distortion Metrics and a need for Perceptual Distortion Metric at a (Sub) Macroblock Level

Yetish G. Joshi, Purav Shah, Jonathan Loo and Shahedur Rahman
Computer & Communications Engineering Department,
School of Science and Technology, Middlesex University, The Burroughs, London NW4 4BT
{y.joshi, p.shah, j.loo, s.rahman}@mdx.ac.uk

***Abstract*—Within a video encoder the distortion metric performs an Image Quality Assessment (IQA). However, to exploit perceptual redundancy to lower the convex hull of the Rate-Distortion (R-D) curve, a Perceptual Distortion Metric (PDM) modelling of the Human Visual System (HVS) should be used. Since block-based video encoders like H.264/AVC operate at the Sub-Macroblock (Sub-MB) level, there exists a need to produce a locally operating PDM. A locally operating PDM must meet the requirements of Standard Traditional Distortion Metrics (STDMs), in that it must satisfy the Triangle Equality Rule ($\trianglelefteq$). Hence, this paper presents a review of STDMs of SSE, SAD and SATD against the perceptual IQA of Structural Similarity (SSIM) at the Sub-MB level. Furthermore, this paper illustrates the Universal Bounded Region (UBR) by block size that supports the triangle equality rule ($\trianglelefteq$) within the Sub-MB level, between SSIM and STDMs like SATD at the prediction stage.**

I. Introduction

Encoders such as MPEG4/AVC (H.264/AVC) and more recently H.265 are deemed as block-based video encoders [1], [2], as they select a prediction mode for a given block with the minimum of pixel difference - residue. This is extended when inter coding is considered as the grouping of (Sub) Macroblocks (Sub-MB) with the minimum amount of motion vectors for the least amount of distortion. Therefore, majority of the block can be represented with signalling and quantisation of the residual pixel difference. This is represented by the Rate-Distortion (R-D) curve in equation (1) in [3], where the effects of lambda (λ) to maintain a given bit rate (R) as part of Rate Control must be assessed by the distortion metric (D).

$$J_{min\,energy} = \lambda_{quant} \times R_{bit\,rate} + D_{dist\,metric} \qquad (1)$$

Hence, the quantisation benefit of lowering the bit rate must be factored with any cost increase in the distortion measured, leading to the search for $J_{min\,energy}$, which can be considered to be an optimum point of operation for the encoder. In particular, the role of the distortion metric is significant within the R-D curve, described in [4] as a convex hull. In the context of the front-end of the encoder, the distortion metric is used in three main areas; selection of the prediction modes, choosing various block sizes during mode decision and assessing the level of activity for the incoming MB when taking Rate Control into account. This process can be illustrated in figure 4 as stages '1' (Distortion Metric - red box), '2' (Mode Decision - green box) and '3' (Rate Control based on [5] - blue dashed outline) respectively.

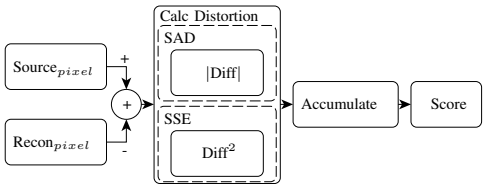


Fig. 1. Distortion Metric Operation with SAD or SSE IQA

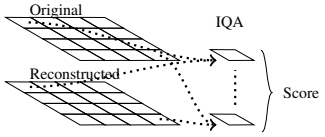


Fig. 2. Standard Traditional Distortion Metric (STDM)'s based IQA

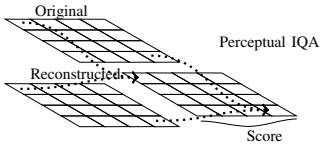


Fig. 3. Perceptual IQA

However, it was discussed in [4], [6] that distortion metrics within video encoders should ideally be based upon the Human Visual System (HVS), though due to reasons of complexity and lack of tractable scoring HVS solutions, they have not been integrated at the block-base level of a video encoder. Instead, Standard Traditional Distortion Metrics (STDMs), such as Sum of Square Errors (SSE) and Sum of Absolute Difference (SAD) are used at the block-base level. These STDMs are simple to operate and tractable, where every pixel difference is uniformly accumulated towards an overall distortion score as shown in figure 1.

An advantage of modelling the HVS is perceptual sensitivity

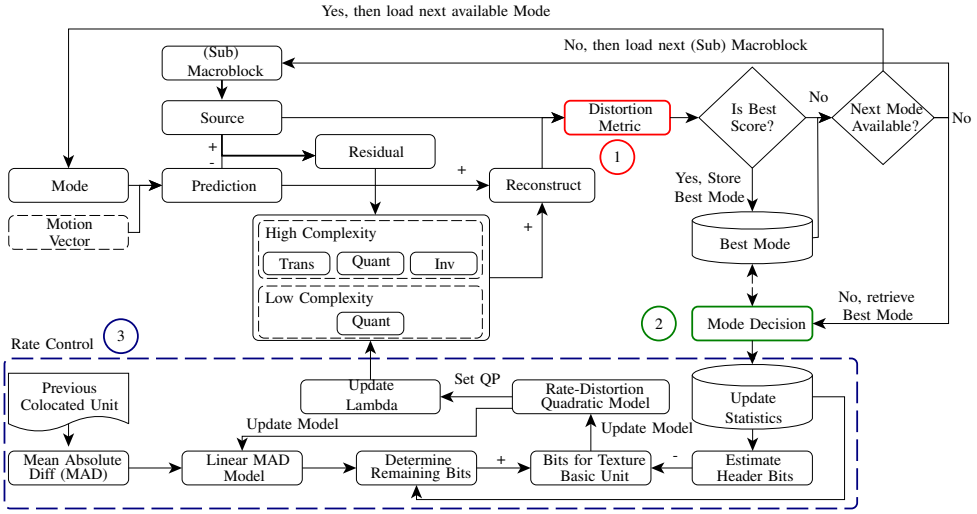


Fig. 4. Front-End Block Based Video Encoder System Level Overview with Rate Control

of relative lighting conditions and structural information can be evaluated compared to absolute pixel difference accumulation of STDMs. This is illustrated in figures 2 and 3 and reinforced in [7]. In terms of a perceptual based modelling, the convex hull can be closer to the origin as distortion is non-uniformly weighted like Just Noticeable Distortion (JND) [8] and Contrast Sensitivity Function (CSF) [9].

JND and CSF perceptual based models reflect the nature of HVS's sensitivity to varying lighting conditions. In these models, least sensitivity is applied to darker regions where objects or texture can be less distinguishable [7]. When edges are visible to the HVS, cognitive sensitivity allows for objects to be recognised by their structural information [10]. Therefore, the HVS relies upon structural information based upon relative lighting conditions to recognise and track objects. Compared to STDMs, where equal weighting is provided, perceptual Image Quality Assessment (IQA) identifies perceptual clues worth retaining and perceptual redundancy that can be exploited for better bit budget utilisation.

However, while these early models of JND and CSF showed the promise to distinguish on perceptual terms, they are complex to implement and operate at the frame level. This computational burden of early models motivated a second generation of application specific perceptual models, primarily for perceptual based coding in video-calling application [11] [12]. Here they focused on reducing the computational complexity by simplifying aspects of these perceptual based models and combining other perceptual based models such as edge detection to produce a multi-HVS perceptual model.

While neither of these application based models replaced the distortion metric, they highlighted the need to do so. The current third generation of HVS modelling took the initiative to consider this direction of modifying the distortion metric and replace it with a multi-HVS based model. This was attempted in [13], [14], however, a perceptual-based model faces the challenge of being implemented within the encoder workflow as a low processing envelope as well as operating as a locally independent operations as discussed in [6]. This has not been successfully achieved to date.

II. A Way Towards Perceptual IQA - Structured Similarity (SSIM)

Structural Similarity (SSIM) [15], a low complexity perceptual Image Quality Assessment (IQA) that takes into account the structural information based on relative lighting conditions and is described in equation (2),

$$SSIM(org,rec)=\frac{(2\mu_{org}\mu_{rec}+C_1)\times(2\sigma_{org,rec}+C_2)}{(\mu_{org}^2+\mu_{rec}^2+C_1)\times(\sigma_{org}^2+\sigma_{rec}^2+C_2)} \quad (2)$$

where, μ_{org} and μ_{rec} represent the mean of the original image block and reconstructed image block, $\sigma_{org}^2$and σ_{rec}^2 are the standard deviations respectively, $\sigma_{org,rec}$ is the covariance, C_1 and C_2 are constants which are calculated based upon the bit depth to stabilise the equation. An extensive study discussed in [16] showed a range of perceptual based IQA's available (including variations of SSIM) being tested and it was concluded that SSIM performed well whilst offering a low processing overhead.

Compared to STDM, SSIM does not support the Triangle Equality Rule ($\trianglelefteq$) natively. In terms of the video encoding, the triangle equality rule is where an image triplet of original, predicted and difference are considered; the distortion score of each pair should be such that the distortion score of one should equate to the summation of the other two sides [17].

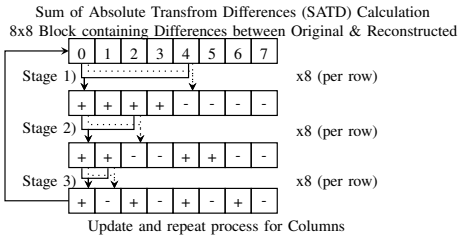


Fig. 5. SATD Operational Block Diagram.

In [13], SSIM has been scaled using logarithmic functions within the distortion metric space of MSE in order for SSIM to support the triangle equality rule ($\trianglelefteq$) and this approach only works at the Group of Picture (GOP) level and being unable to adapt to the local changes. Furthermore, in [13], a perceptual measure is scaled into the distortion space of MSE, where the distortion scale can potentially be wide. Typically, for an 8-bit Luma pixel depth this means a theoretical maximum of 255^2.

The applications of perceptual IQA limited to GOP and frame level, restricts the effect on bringing R-D curve closer to the origin. Therefore, the process of mapping perceptual IQA on non-Perceptual Distortion Metric (PDM) should be extended to the Sub-MB level to reflect the operation of a block based video encoder. In addition, non-PDM should be evaluated in terms of their complexity and potential range of distortion scores for a low processing overhead and limited range of values respectively.

Thus, this paper will investigate whether a SSIM based PDM can exist at the Sub-MB level. This will be done by assessing Sub-MBs simultaneously under SSIM and STDM, evaluating whether SSIM operates within a closed distortion metric space of STDMs. A closed distortion metric space will indicate that SSIM can be scaled to satisfy the triangle equality rule ($\trianglelefteq$). Hence, a future block-based encoder using a scaled-SSIM-PDM can achieve a lower convex hull R-D curve.

III. SSIM within the Distortion Metric Space of STDMs

The independent dimensionless pixel level evaluation of STDMs such as SSE and SAD are scalable, unaffected by adjacent pixel differences. Therefore, STDMs are unable to appreciate the significance of inherent visual clues like structure or texture within a Sub-MB. Another distortion metric to consider is SATD, which utilises the Hadamard Transform and is designed to be processor friendly as shown in figure 5, already used in H.264/AVC's back-end [18].

Unlike other STDMs that have a high dependency on computational loops, SATD is an efficient alternative [18] as it utilises shifts, addition and subtraction. However, these STDMs are weightless metric, meaning each difference is treated equally. In perceptual IQA, it has not yet been discussed whether the amount of neighbouring pixels would affect the performance at the Sub-MB level. It is important to understand that SSIM is an averaging of a series of sliding windows of local SSIM's and hence, it is about determining the size of the block and the amount of overlap between blocks [19]. It was shown in [15] that a block size of 8x8 pixels is recommended to provide a stable result and a greater degree of overlap between blocks would provide an accurate SSIM result. SSIM also supports the small 4x4 block size of Sub-MB [20].

Having a PDM will influence both intra and inter blocks across each of the stages as shown in figure 4 as part of the Perceptual Framework design. To minimise the processing load, the SSIM window size will be equal to the Sub-MB size. This approach can be extended if needed, to produced a more accurate SSIM, by using smaller window sizes and overlapping windows at the expense of additional processing.

IV. An Investigation into Perceptual IQA at the Prediction Stage with Sub-Macroblocks

In order to lower the convex hull of the R-D curve, it is necessary to have a PDM working at the Sub-MB level. This paper introduces SSIM at the prediction level in order to assess its feasibility to work at the Sub-MB level against STDMs. The results presented in figure 6 and figure 7 have been extracted from the JM18.4 H.264/AVC [18] encoder, which has been modified to incorporate SSIM at the prediction stage with SSIM window size equal to the Sub-MB size. The default configuration file for JM18.4 complies predominately with the recommendations set in [21] with only minor changes required.

The video sequence used for these tests is chosen as the Foreman video, with QCIF resolution of 176×144 pixels and consists of three frames. At the prediction stage in the Sub-MB level, the iterative operations result in 900k and 700k samples captured for the 4x4 and 8x8 block respectively across both the inter frames in the test video sequence.

The test results in Figure 8 were obtained using higher CIF resolution based video sequences of varying content with only 4x4 and 8x8 inter blocks considered to validate the earlier findings of figure 7.

Focusing on the Intra graphs as shown in figure 6, a concentration of samples can be described close to the origin highlighting the statistical perceptual similarity of intra prediction. In 4x4, a broad range of 1-SSIM values exist for a limited range of STDM score, suggesting that SSIM evaluates with greater sensitivity when in smaller block sizes.

In terms of the 8x8 Intra graphs, the results seem more narrow, usually with most samples concentrated within the first 0.25 of (1-SSIM) range. This suggests that 8x8 does encounter predictions that are favourable for SSIM than STDMs.

The results for the Inter blocks as shown in figure 7 have improved definition of the distortion metric space than of Intra. This is because RDO is enabled leading to permutations of mode predictions and motion vector predictions being considered. As such, 4x4 blocks of Inter extend 1-SSIM to 0.75, where as in the 8x8 configuration, the shape of the distortion metric space is beginning to appear with trails of samples extending along the x-axis beyond 1. The

samples where (1-SSIM) is <1, more perceptual information is stored, conversely; samples >1 have high amounts of blocking artefacts making it perceptually unrecognisable.

Upon analysing the distortion score ranges, it is found that SSE has the highest range of 125k for 4x4 and 250k for 8x8. Theoretically, this could be as high as 1 million and 4 million respectively in this case which is highly unlikely. With regards to the scales recorded against the theoretical highs, this represents as a fraction 1/8th of 4x4 and 1/16th of 8x8 SSE's distortion metric space. For SAD and SATD, they cover a greater proportion of the maximum possible scores, $\approx 3/10$ and $1/4$ for 4x4 and 8x8 respectively. Knowing that SAD and SATD have a smaller dynamic range and the samples cover a larger proportion of the distortion metric space, SATD allows for any potential model to be mapped with greater coverage.

Overall, analysing figures 6 and 7 by block size, indicates that two scaled-SSIM models by block size are required to produce a scaled-SSIM-PDM, as the graphs illustrate Intra to be a limited version of the Inter.

As SATD is the preferred distortion metric in [18] due to its processor friendly operations, and in order to validate the relationship of 1-SSIM and STDM, further results were gathered. Following the results presented in figures 6 and 7, it has been possible to replicate the relationship of perceptual IQA vs. non-PDM using SSIM and SATD respectively with higher resolution video sequences. This is shown in figure 8, where CIF resolution is used with the number of samples gathered in excess of four million. The overall shape is the same as seen earlier with Foreman (QCIF), though depending on the nature of the video the scores differ. This shows that a scaled-SSIM-PDM can exist within the STDM distortion metric space that satisfies the triangle equality rule ($\trianglelefteq$).

Therefore, these findings are significant as it reflects that a Universal Bounded Region (UBR) by block size at the Sub-MB level exists, irrespective of video resolution or the type of video sequence. This supports the case for the SSIM is mapped against an STDM space.

V. Conclusions and Future Work

HVS offers the ability to assess perceptually significant and redundant information. To effectively implement a perceptual based HVS model at the encoder system level, it requires to be integrated at the Sub-Macroblock level. However, the triangle equality rule ($\trianglelefteq$) inhibits perceptual IQA such as SSIM from being adopted at the Sub-MB level. This paper has presented the evidence that a Perceptual IQA - SSIM, at a Sub-MB level has a relationship with STDMs. This was further confirmed by higher resolution video, illustrating that this relationship is independent of the video resolution and type of sequence. Hence, a Perceptual Distortion Metric (PDM) can be modelled by scaling SSIM within what is labelled as the Universal Bounded Region (UBR) by block size, thus satisfying the triangle equality rule ($\trianglelefteq$).

Furthermore, a Perceptual Framework can be designed around PDM to affect the highlighted regions of distortion metric, mode decision and rate-control as shown in figure 4.

References

[1] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC Video Coding Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[2] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649 –1668, dec. 2012.

[3] H. Everett III, "Generalised Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources," *Operations Research*, vol. 11, no. 3, pp. 399 – 417, 1963.

[4] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 23–50, 1998.

[5] Z. Li, W. Gao, F. Pan, S. Ma, K. Lim, G. Feng, X. Lin, S. Rahardja, H. Lu, and Y. Lu, "Adaptive Rate Control for H.264," *Journal of Visual Communication and Image Representation*, vol. 17, no. 2, pp. 376 – 406, 2006.

[6] G. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, 1998.

[7] H. . R. . Wu and K. . R. . Rao, Eds., *Digital Video Image Quality and Perceptual Coding*. CRC Press, 2005.

[8] C.-H. Chou and Y.-C. Li, "A Perceptually Tuned Subband Image Coder Based on the Measure of Just-Noticeable-Distortion Profile," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 6, pp. 467–476, 1995.

[9] J. Yogeshwar and R. J. Mammone, "A New Perceptual Model for Video Sequence Encoding," in *Proc. Conf. th Int Pattern Recognition*, 1990, pp. 188–193.

[10] S. J. S. Robin E. N. Horne, Ed., *The Colour Image Processing Handbook*. Springer Berlin / Heidelberg, 1998.

[11] R. Jin and J. Chen, "The Coding Rate Control of Consistent Perceptual Video Quality in H.264 ROI," in *Proc. Int. Symp. Computer Network and Multimedia Technology CNMT 2009*, 2009, pp. 1–4.

[12] X. K. Yang, W. S. Lin, Z. K. Lu, X. Lin, S. Rahardja, E. P. Ong, and S. S. Yao, "Local Visual Perceptual Clues and its use in Videophone Rate Control," in *Proc. Int. Symp. Circuits and Systems ISCAS '04*, vol. 3, 2004.

[13] Y.-H. Huang, T.-S. Ou, P.-Y. Su, and H. Chen, "Perceptual Rate-Distortion Optimization using Structural Similarity Index as Quality Metric," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1614 –1624, 2010.

[14] A. Bhat, I. Richardson, and S. Kannangara, "A New Perceptual Quality Metric for Compressed Video Based on Mean Squared Error," *Signal Processing: Image Communication*, vol. 25, no. 8, pp. 588–596, 2010.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[16] W. Lin and C.-C. J. Kuo, "Perceptual Visual Quality Metrics: A Survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297 – 312, 2011.

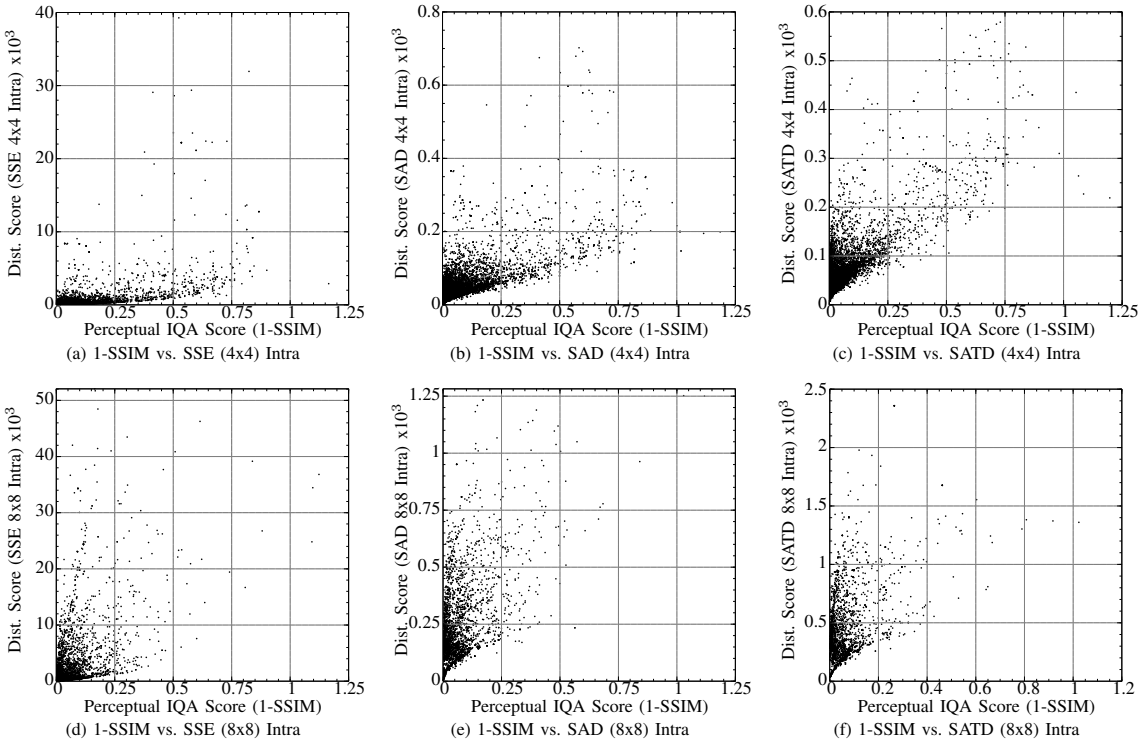
[17] T. Richter, "SSIM as Global Quality Metric: A Differential Geometry View," in *Proc. Third Int Quality of Multimedia Experience (QoMEX) Workshop*, 2011, pp. 189–194.

[18] K. Sühring. H.264/AVC Reference Software JM. [Online]. Available: http://iphome.hhi.de/suehring/tml/

[19] D. Brunet, "A Study of the Structural Similarity Image Quality Measure with Applications to Image Processing," Ph.D. dissertation, University of Waterloo, Sept 2012. [Online]. Available: http://hdl.handle.net/10012/6982

[20] A. Brooks, X. Zhao, and T. Pappas, "Structural Similarity Quality Metrics in a Coding Context: Exploring the Space of Realistic Distortions," *IEEE Transactions on Image Processing*, vol. 17, no. 8, pp. 1261 –1273, aug. 2008.

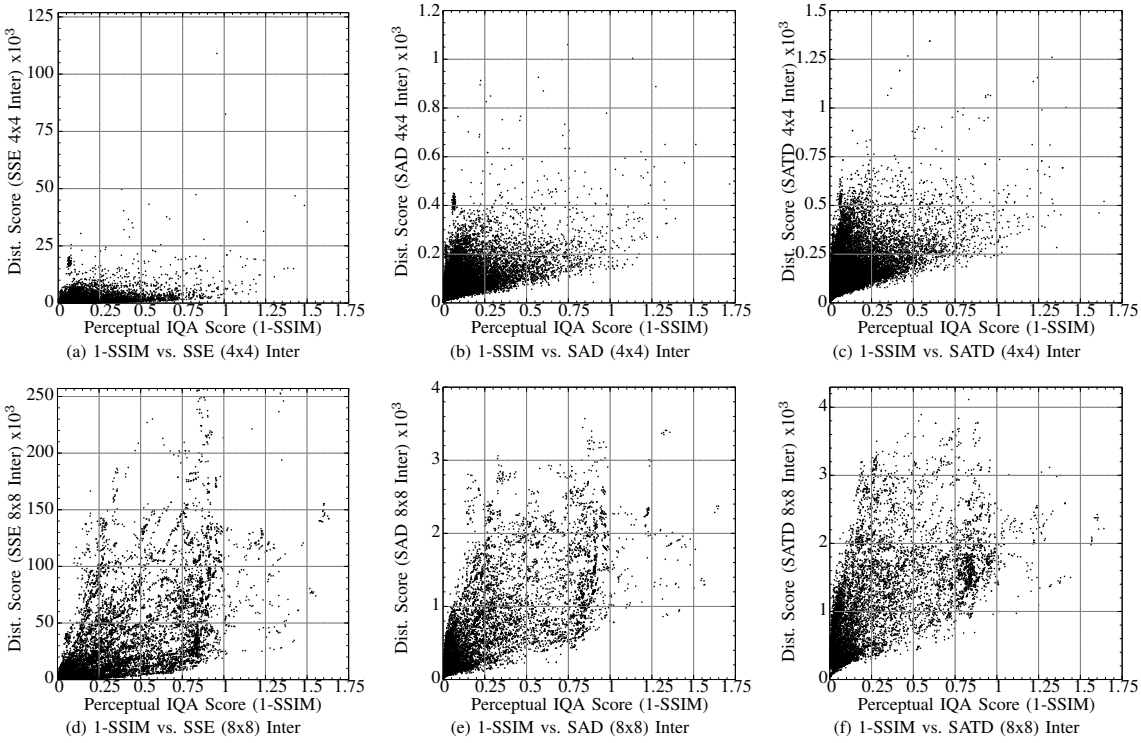
[21] S. G. Tan T.K. and W. T., "(VCEG-AJ10r1) Recommended Simulation Common Conditions for Coding Efficiency Experiments Revision 4." ITU-T SC16/Q6, 36th VCEG Meeting, San Diego, USA, 8th - 10th Oct., 2008, 2008.



(a) 1-SSIM vs. SSE (4x4) Intra (b) 1-SSIM vs. SAD (4x4) Intra (c) 1-SSIM vs. SATD (4x4) Intra

(d) 1-SSIM vs. SSE (8x8) Intra (e) 1-SSIM vs. SAD (8x8) Intra (f) 1-SSIM vs. SATD (8x8) Intra

Fig. 6. Perceptual Image Quality Assessment (IQA) vs. Distortion Metric from 4x4 and 8x8 Intra Blocks. Structural Similarity (SSIM), plotted against Sum of Square Errors (SSE), Sum of Absolute Difference (SAD) and Sum of Absolute Transform Difference (SATD).



(a) 1-SSIM vs. SSE (4x4) Inter (b) 1-SSIM vs. SAD (4x4) Inter (c) 1-SSIM vs. SATD (4x4) Inter

(d) 1-SSIM vs. SSE (8x8) Inter (e) 1-SSIM vs. SAD (8x8) Inter (f) 1-SSIM vs. SATD (8x8) Inter

Fig. 7. Perceptual Image Quality Assessment (IQA) vs. Distortion Metric from 4x4 and 8x8 Inter Blocks. Structural Similarity (SSIM), plotted against Sum

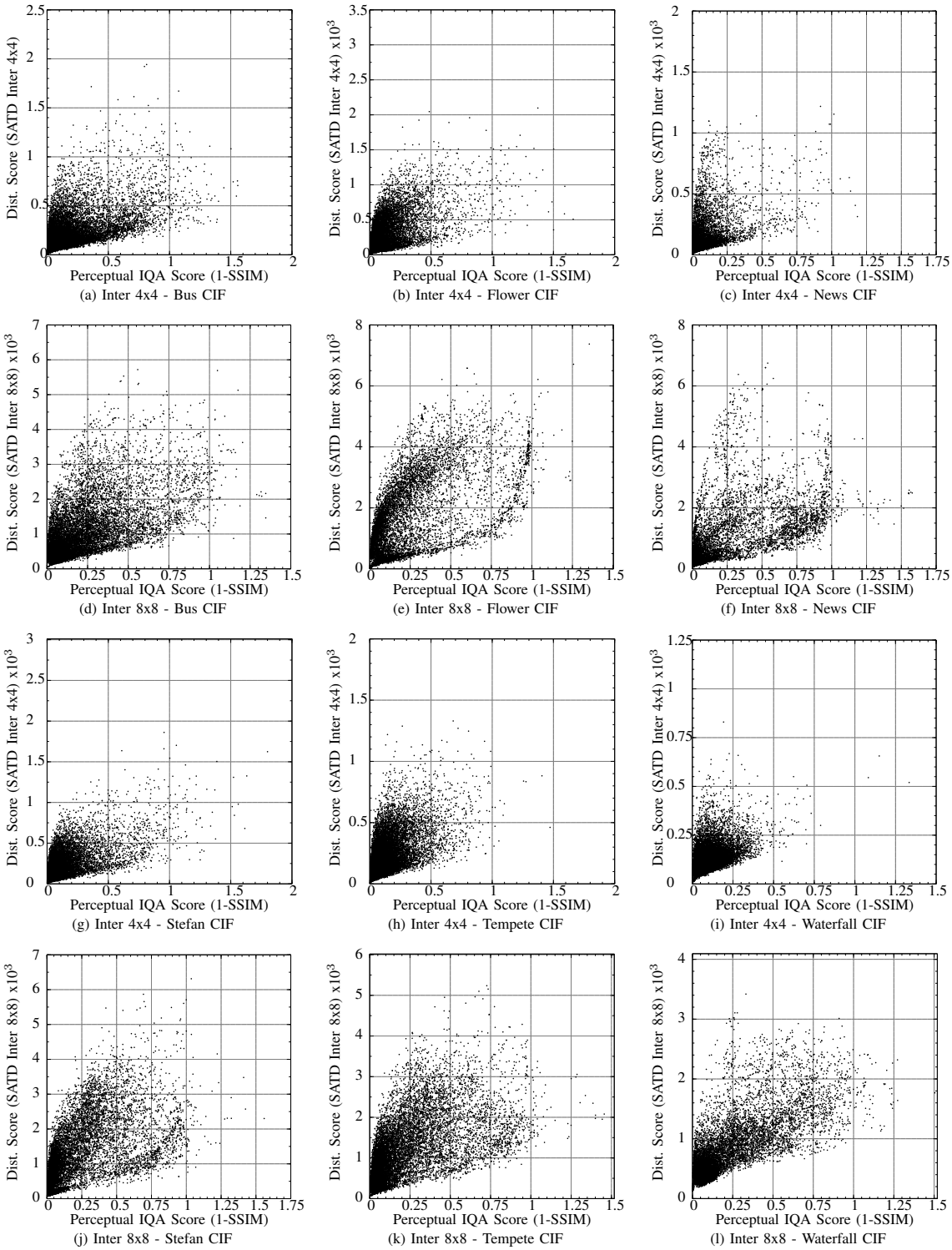


Fig. 8. Perceptual Image Quality Assessment (IQA) vs. Distortion Metric from 4x4 and 8x8 Inter Blocks. Structural Similarity (SSIM), plotted against Sum of Absolute Transform Difference (SATD) for CIF Video Sequences (Bus, Flower, News, Stefan, Tempete and Waterfall). First three frames were used and over four million samples gathered. Graphs show thinned results by a factor of 250.

A Novel Low Complexity Local Hybrid Pseudo-SSIM-SATD Distortion Metric Towards Perceptual Rate Control

Yetish G. Joshi*, Jonathan Loo*, Purav Shah*, Shahedur Rahman* and Yoong Choon Chang †
* Computer and Communications Engineering, School of Science and Technology, Middlesex University, London NW4 4BT
{y.joshi, j.loo, p.shah, s.rahman}@mdx.ac.uk
† Faculty of Engineering, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia
ycchang@mmu.edu.my

***Abstract*—The front-end block-based video encoder applies an Image Quality Assessment (IQA) as part of the distortion metric. Typically, the distortion metric applies uniform weighting for the absolute differences within a Sub-Macroblock (Sub-MB) at any given time. As video is predominately designed for Humans, the distortion metric should reflect the Human Visual System (HVS). Thus, a perceptual distortion metric (PDM), will lower the convex hull of the Rate-Distortion (R-D) curve towards the origin, by removing perceptual redundancy and retaining perceptual clues. Structured Similarity (SSIM), a perceptual IQA, has been adapted via logarithmic functions to measure distortion, however, it is restricted to the Group of Picture level and hence unable to adapt to the local Sub-MB changes. This paper proposes a Local Hybrid Pseudo-SSIM-SATD (LHPSS) Distortion Metric, operating at the Sub-MB level and satisfying the Triangle Equality Rule (◁). A detailed discussion of LHPSS's Psuedo-SSIM model will illustrate how SSIM can be perceptually scaled within the distortion metric space of SATD using non-logarithmic functions. Results of HD video encoded across different QPs will be presented showing the competitive bit usage under IbBbBbBbP prediction structure for similar image quality. Finally, the mode decision choices superimposed on the Intra frame will illustrate that LHPSS lowers the R-D curve as homogeneous regions are represented with larger block size.**

I. Introduction

The role of the front-end block-based video encoder is to select the prediction representing the most amount of pixel image block as signalling, for the least amount of distortion for the quantised residue. This is reflected by the Rate-Distortion (R-D) curve in equation (1) within [1]. Here, lambda (λ) applies quantisation to maintain a given bit rate (R), while its effects are assessed by the distortion metric (D). This can extend along various stages of the encoder [2], searching for $J_{min\,energy}$, the optimum point of operation along the convex hull of the R-D curve for the encoder as discussed in [3].

$$J_{min\,energy} = \lambda_{quant} \times R_{bit\,rate} + D_{dist\,metric} \qquad (1)$$

The benefits of having a distortion metric based upon the HVS (perceptual model) can bring the convex hull closer to the origin, thus lowering the bit-rate required to achieve similar Image Quality (IQ) [3], [4]. However, perceptual models can be computationally high and their perceptual distortion scores difficult to quantify [4], hence, the use of low complex tractable solutions [3], which support the Triangle Equality Rule (◁) [5].

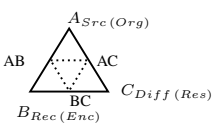


Fig. 1. Triangle Equality

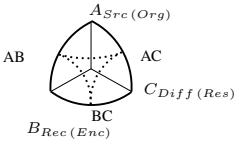


Fig. 2. Geodesic Triangle Equality

The triangle equality rule (◁) in terms of the distortion metric is where for an image triplet of original, predicted and difference; the distortion score of each pair should be such that the distortion score of one should equate to the summation of the other two distortion scores as shown in figure 1 [5].

Different perceptual based models have achieved the goal of lowering the convex hull towards the origin, such as Just Noticeable Distortion (JND) [6]. JND is highly complex and considers relative lighting conditions, an aspect of HVS. Structural Similarity (SSIM) [7], a low complexity perceptual Image Quality Assessment (IQA) that takes into account the structural information based on relative lighting conditions and is described in equation (2),

$$SSIM(org, rec) = \frac{(2\mu_{org}\mu_{rec}+C_1)\times(2\sigma_{org,rec}+C_2)}{(\mu_{org}^2+\mu_{rec}^2+C_1)\times(\sigma_{org}^2+\sigma_{rec}^2+C_2)} \qquad (2)$$

where, μ_{org} and μ_{rec} represent the mean of the original image block and reconstructed image block, $\sigma_{org}^2$and σ_{rec}^2 are the standard deviations respectively, $\sigma_{org,rec}$ is the covariance, C_1 and C_2 are constants which are calculated based upon the bit depth to stabilise the equation.

It was explained in [5], that a true distortion metric supports the triangle equality rule (◁), suggesting that SSIM should support a Geodesic Triangle Equality, as shown in figure 2, over a curved space. Hence, non-linear equations should be applied to SSIM to scale it such that it satisfies the triangle equality rule (◁).

SSIM is a perceptual IQA, but cannot be natively used as a distortion metric as it does not support the triangle equality

rule ($\trianglelefteq$). The efforts of adapting SSIM to operate as a pseudo-distortion metric have been achieved in [8] by use of logarithmic functions, however this approach is limited to the Group of Pictures (GOP) level. Furthermore, it does not meet the goals of low complexity and variability [3]. Variability distinguishes two similar results by their scores, which when considered at the Sub-Macroblock (Sub-MB) level at the Prediction and Mode Decision stages, is crucial. Hence, the accuracy and coverage of a perceptual distortion metric (PDM) must be sufficient to achieve this for the selection of prediction modes and block sizes. The scaled-SSIM-PDM is the evidence backed concept of representing SSIM values within the Standard Traditional Distortion Metrics (STDM) space, proposed in [2]. Though no means of realising this concept was shown, the work highlighted that logarithmic functions should be avoided. Similar to IQA, perceptual vs. non-perceptual, a scaled-SSIM-PDM vs. a non-PDM will differ by the ordering of scores, allowing for certain types of distortions over others [7]. This can be extended at the local level, where the scaling can be adapted according to the perceptual or bit-budget conditions at that given time. Therefore, equation (1) can be re-written as equation (3), where kappa ($\boldsymbol{\kappa}$) represents adapting the scaling of the PDM towards a perceptual rate control (PRC). Rather than adjusting λ to regulate the bit-budget, $\boldsymbol{\kappa}$ can influence the PDM based upon the perceptual significance of the incoming MB [2].

$$J_{min\,energy} = \lambda_{quant} \times R_{bit\,rate} + \boldsymbol{\kappa} \times D_{dist\,metric} \quad (3)$$

Hence, this paper will demostrate SSIM scaled in the distortion metric space of SATD at the Sub-MB level, avoiding logarithmic functions. This will be shown in the form of Pseudo-SSIM and as part of Local Hybrid Pseudo-SSIM-SATD (LHPSS) Distortion Metric which falls back to SATD when Pseudo-SSIM is out of scope.

The paper is divided up as follows, an explanation of how covariance can aid in ordering samples that occupy the same SSIM value, thus allowing SSIM to be scaled and to meet the triangle equality rule ($\trianglelefteq$). However, covariance should be perceptually compared to a HVS model like Just Noticeable Distortion (JND) [6] to assess how it perceptually evaluates an image. Then, a flowchart and operational block diagram will illustrate the Local Hybrid Pseudo-SSIM-SATD (LHPSS) Distortion Metric and Pseudo-SSIM operations in figure 6 and section IV respectively. Finally, the results of the implemented LHPSS, in terms of table of results and Intra frames with mode decision superimposed shown in table I and figure 8 respectively. Please note that the scaling values of Pseudo-SSIM are provided in tables II to XIV.

II. Ordering of SSIM within an existing Distortion Metric Space

It is shown in [5] that SSIM must be non-negative, symmetrical and fulfil the triangle equality rule ($\trianglelefteq$). These first two conditions are met by presenting SSIM in the form of (1-SSIM), a method adopted in [8] and explained in [5]. To support the triangle equality rule ($\trianglelefteq$), the findings in [2] of a universal bounded region (UBR) by block size is extended in this paper by stating how they may be ordered.

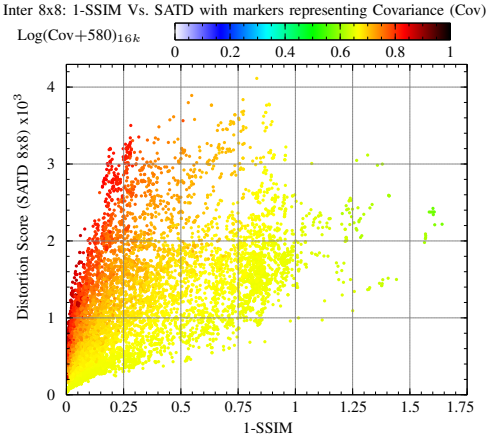


Fig. 3. SSIM vs. SATD (8x8) with Covariance.

The graph in figure 3 illustrates SSIM samples taken along side SATD samples at the Prediction level, with the markers coloured by their covariance (Cov) value. Covariance is a component of SSIM, calculated between the original and reconstructed block as shown in equation (2). In terms of 8-bit Grey-scale, Luma, the theoretical range for Covariance is $\pm$16k. In the graph figure 3, the actual range observed were between -334 and 4731, hence the covariance value has been shifted by $+580$, ($C_2 \times 10$), so that negative covariance values can be illustrated on the same graph. From figure 3, it shows how samples that occupy the same SSIM value can be distinguished by their covariance value. Thus, figure 3 provides insight of how the concept of a scaled-SSIM-PDM in [2] can be achieved within distortion metric of SATD and thus satisfying the triangle equality rule ($\trianglelefteq$).

For reasons of time and simplicity, the modelling of figure 3 in terms of Pseudo-SSIM will be bounded, where $0 \leq$ Cov $<$ 8000 and where 1$-$SSIM $<$1. This should cover the majority of samples as shown in [2], however when out of scope of Psuedo-SSIM it will falling back to SATD .

Pseudo-SSIM model will enable prediction and mode decision to assess in terms of the perceptual score at the Sub-MB level. This can be further extended in the form of perceptual rate control as described in equation (3).

III. Covariance Map Analysis

In order to assess the perceptual nature of covariance, a covariance heatmap based upon raw Luma values was produced using a spreadsheet. This was based upon raw Luma values of original and reconstructed intra frames, the covariance range was shifted by three to ensure full coverage. Figure 4 represents the values of covariance in a heat map format with a scale of $Log_{8000}(Cov + 3)$.

The Covariance Map illustrates flat regions with low covariance and where edges or boundaries exist, represented by high covariance. To compare the perceptual covariance

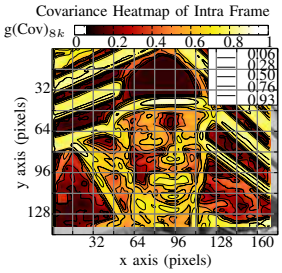


Fig. 4. Covariance Heatmap of Intra Frame (Foreman frame 0 QCIF).

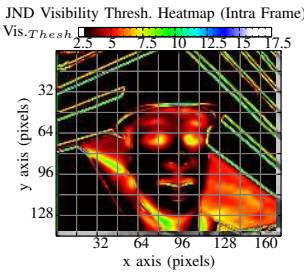


Fig. 5. Just Noticeable Distortion (JND) Visibility Threshold of Intra Frame.

heatmap in HVS terms, another heatmap representing JND using [6] was produced based upon raw luma values and a spreadsheet. The JND heatmap is shown figure 5 illustrates the visibility threshold to be lowest in the homogeneous regions, i.e. the bottom left hand corner, the helmet and upon the panelling. These have the darkest region, where the sensitivity to Luma differences are high and most noticeable. Overall, analysing figure 4 and figure 5 shows that covariance makes a reasonable approximation of JND. This justifies covariance as a perceptual means of scaling of SSIM values. Furthermore, this understanding of interpreting local covariance values of original and reconstructed images can be used to interpret the graphs shown in [2].

IV. Local Hybrid Pseudo-SSIM-SATD (LHPSS) Distortion Metric

Within the encoding process, the Local Hybrid Pseudo SSIM-SATD (LHPSS) model will affect both the intra and inter blocks at the mode and prediction levels and can be extended to Rate Control as part of the PRC model [2].

As Pseudo-SSIM has been defined as where $1-\text{SSIM}<1$ and $0\leq\text{Cov}<8000$, it must work in a Hybrid form along side the distortion metric it mimics, SATD, as a fall-back to ensure full coverage. Figure 6 represents a flowchart of the LHPSS Distortion Metric. Within the flowchart, the Absolute Mean Difference ($|\mu_{O-R}|$) and Covariance are used to provide variableness and scale SSIM respectively. Variableness between samples is crucial for the encoder to distinguish between similar samples [3], especially at the Sub-MB level where the likelihood of prediction modes sharing the same SSIM score is high [2].

The workflow of LHPSS as shown in figure 6 uses SSIM's own components to scale Pseudo-SSIM. The scaling is performed using linear equations to ensure processor friendly operations. Compared to [8], LHPSS can operate locally without using logarithmic functions and without re-quantising to produce a temporal relative distortion scale. As LHPSS operates in the distortion metric space of SATD it satisfies the triangle equality rule ($\trianglelefteq$). Therefore, in [8] the perceptual model must be actively updated on a key frame basis as its relative nature limits the scope to the GOP level or when there is high activity. Hence, in [8] it does not adapt to the local conditions like an STDM. LHPSS distortion metric

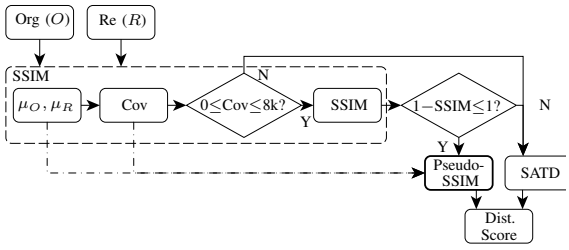


Fig. 6. Flowchart of Hybrid Pseudo-SSIM-SATD Distortion Metric.

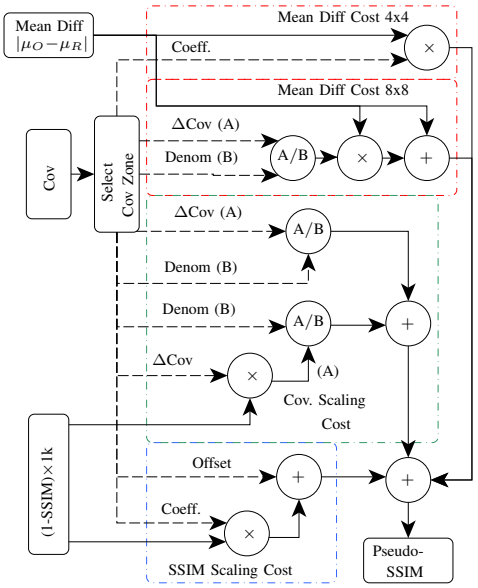


Fig. 7. Operational Block Diagram of PseudoSSIM

addresses these issues related to [8] by providing a general solution based upon modelling the UBR by block size [2]. This enables LHPSS to be integrated at the Sub-MB prediction and mode decision stages, thus awarding a distortion score based on the local perceptual significance. Furthermore, Pseudo-SSIM within LHPSS can be extended to dynamically adapt as described in equation (3)

From an operational point of view, Pseudo-SSIM can be seen to be competitive against STDMs. While an STDM would calculate the differences between a Original and Reconstructed block before determining the score, SSIM can be performed immediately before Pseudo-SSIM scales the SSIM value. However, where the SSIM or covariance values are out of range, it must then fall back to use SATD, an STDM. Therefore, for those values that fall out of scope of Pseudo-SSIM, the encoding process takes longer, thus requiring additional processing time.

V. Working of Pseudo-SSIM within LHPSS

The samples gathered to produce the scaled-SSIM-PDM model of Pseudo-SSIM were taken at the prediction level [2] based upon the data gathered from the first three frames of Foreman video sequence of QCIF resolution and reflect the large proportion low covariance regions shown in figure 4.

The internal processes involved to calculate Pseudo-SSIM are shown in figure 7. They comprise of three main parts, the Mean Difference Cost, the Covariance Scaling Cost and the SSIM Scaling Cost and together they sum up to output a Pseudo-SSIM score. There are six zones, which are set depending upon the covariance score, where ΔCov is the relative covariance that is subsequently processed. The weightings per covariance zone and SSIM profiles for Pseudo-SSIM are shown in tables II to XIV.

Within the Mean Difference Cost the pathways are specific to the block size. Specifically, the 8x8 Model, the mean difference can be high which will lead to addition processing of the mean difference towards the final value of the Mean Difference Cost. Otherwise, the Covariance Scaling Cost and SSIM Scaling Cost for both 4x4 and 8x8 block sizes are the same with only the values that differ. The values set for coefficients, denominator and offset within Pseudo-SSIM have been set to be either binary friendly or when applied implemented with as shifts, additions or subtractions. This has been possible by working with integers. The SSIM score is initially converted to $(1-\text{SSIM})\times 1000$ and covariance has had its respective covariance zones threshold subtracted and labelled as ΔCov. Thus, only when two unknowns at designed time are multiplied or SSIM is converted to $(1-\text{SSIM})\times 1000$ does a multiplication take place. Therefore, 4x4 block will undergo two multiplication operations and an 8x8 block will have three multiplication operations, The purposed of the covariance scaling cost is to distinguish two samples of same the SSIM value by their respective covariance within the distortion metric space of SATD. Hence, the upper part of ΔCov divided by a denominator factor states the position within the given zone. While the lower part of $\Delta\text{Cov}\times(1-\text{SSIM})\times 1000$ reflects how the zones are divided into SSIM bands and so this represents the rate of growth for the given band.

VI. Results

The results in table I show the implementation of the LHPSS at the Prediction and Mode Decision stages, operating as a distortion metric alternative to SATD where conditions are met. The performance results are from the encoder's console and statistics file and further analysis of the mode selection on the intra frame were extracted separately from the encoder.

The default configuration file in JM18.4 MPEG4/AVC [9] was set-up with the recommendations set by [10] and SSIM assessment. A separate modified JM18.4 code base with the LHPSS model implemented was set-up with the same configuration except for LHPSS operating at Motion Estimation (Half and Quarter Pixel) and at Mode Decision Distortion.

The video sequences selected are of HD resolution (1920x1080), 'CrowdRun' and 'sunflower', 50 and 25 frames/second respectively. The tests were run under 'IbBbBbBbP' and 'IPPP' prediction structure [10] across four Quantisation Parame (QP) values of 22, 27, 32 and 37 for QPISlice, with QPPSlice and QPBSlice increamented by 1, i.e. if QPISlice is 22, QPPSlice is 23 and QPBSlice is 24.

IbBbBbBbP					
CrowdRun	**QP22**	**QP27**	**QP32**	**QP37**	**Ave.**
Total Time	19.70%	19.70%	12.67%	21.07%	18.29%
Y-PSNR	-0.35%	-0.55%	-0.65%	-0.57%	-0.53%
Y-SSIM	-0.06%	-0.20%	-0.44%	-0.68%	-0.34%
Total Bits	1.95%	1.38%	0.53%	-0.54%	0.83%
Sunflower	**QP22**	**QP27**	**QP32**	**QP37**	**Ave.**
Total Time	21.16%	21.91%	23.50%	24.81%	22.85%
Y-PSNR	-0.19%	-0.30%	-0.34%	-0.43%	-0.31%
Y-SSIM	-0.03%	-0.07%	-0.16%	-0.39%	-0.16%
Total Bits	0.57%	-1.99%	-5.45%	-9.49%	-4.09%
IPPP					
CrowdRun	**QP22**	**QP27**	**QP32**	**QP37**	**Ave.**
Total Time	-5.51%	4.80%	10.39%	12.31%	5.50%
Y-PSNR	-31.17%	-32.74%	-35.15%	-38.56%	-34.40%
Y-SSIM	-16.96%	-25.68%	-35.48%	-46.88%	-31.25%
Total Bits	-82.97%	-83.33%	-82.34%	-84.37%	-83.25%
Sunflower	**QP22**	**QP27**	**QP32**	**QP37**	**Ave.**
Total Time	13.27%	19.29%	20.00%	23.90%	19.12%
Y-PSNR	-15.10%	-17.51%	-19.29%	-23.63%	-18.88%
Y-SSIM	-2.80%	-4.80%	-7.88%	-13.88%	-7.34%
Total Bits	-68.86%	-65.60%	-50.29%	-24.49%	-52.31%

TABLE I
SUMMARY OF LHPSS RELATIVE VIDEO PERFORMANCE SHOWN AS % DIFFERENCES FOR IBBBBBBP AND IPPP PREDICTION STRUCTURE USING CROWDRUN AND SUNFLOWER 1080P

The run of tests were performed with Rate Distortion Optimisation (RDO) Quantisation (RDOQ) enabled, whereby each prediction mode is assessed by their Distortion Score as well as compressibility, reflecting the need to balance (R-D). The run of tests with RDOQ disabled have not been performed, since the results with RDOQ enabled would occupy a smaller range of SSIM, excluding the extreme cases and thus, increasing the likelihood of having those prediction modes that are closer to the origin of the R-D curve. The tests were performed using a system with an Intel Core i7 CPU 920 processor operating at 2.67GHz and 7GB of RAM.

VII. Encoder Performance

The novel LHPSS model has been implemented within the JM18.4 H.264/AVC Encoder as shown in figure 7.

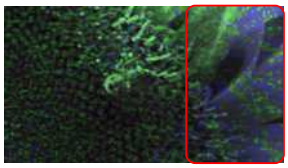
The summary of results presented in table I, shows the performance of SATD and Pseudo-SSIM. The Encoder outputs information pertaining frame bit usage, timings and IQ. These are shown as relative video performance shown as % differences for IbBbBbBbP and IPPP prediction structure using CrowdRun and Sunflower 1080p.

The results for IbBbBbBbP prediction structure illustrate CrowdRun to have an overall bit usage approximately the same as SATD, within $\pm 2\%$, across the range of QP's tested. Similarly, the PSNR and SSIM values remain within 1% difference. Comparing to Sunflower, which is a highly textured video sequence, the bit usage progressively drops under LHPSS as

(a) SATD I4x4:57, I8x8:6408, I16x16:1695

(b) Pseudo-SSIM I4x4:67, I8x8:6161, I16x16:1932

(c) SATD I4x4:2910, I8x8:4333, I16x16:917



(d) Pseudo-SSIM I4x4:2127, I8x8:4880, I16x16:1153

Fig. 8. CrowdRun (top pair) and Sunflower (bottom pair)Frame 1 (Intra) Luma with Highlighted Macroblocks type, Red for Intra4x4, Green for Intra 8x8 and Blue for Intra 16x16.

QP increases, while image quality (IQ) score dropping by less than $\frac{1}{2}\%$. The overall time for encoding the video sequence using the LHPSS model is $+18\%$ and $+23\%$ respectively for CrowdRun and Sunflower.

Under IPPP prediction structure, the overall bit usage dramatically drops by 83% and 52% respectively for CrowdRun and Sunflower; however, this happens at the expense of IQ. For CrowdRun, PSNR using LHPSS reduces by 31% at QP22 and 39% at QP37 when comparted to SATD. For Sunflower, the PSNR is 15% lower at QP22 and 24% lower at QP37 when LHPSS is used. In SSIM terms, the drop in IQ is on average almost a third in CrowdRun and a fifth in Sunflower.

The intra frame with mode selection superimposed on top is shown in figure 8. Here, a greater number of larger block sizes are chosen with LHPSS. In CrowdRun, the top 'sky' region, which is mainly homogeneous, shows a large number of 8x8 blocks (green) when compared to SATD. Also, in the middle front of the CrowdRun, the number of 4x4s used is less under LHPSS. In the Sunflower test video, the number of 16x16 is approximately 14% higher, which are largely concentrated on the petals on the right hand side. Again, this region is homogeneous, thus exploiting perceptual redundancy.

VIII. Conclusion and Future Work

The results presented in this paper have demonstrated that SSIM scaling by using its own component of 'covariance', both satisfies the Triangle Equality Rule ($\trianglelefteq$) and utilises a perceptual means of scaling.

The LHPSS model utilises non-logarithmic functions thus allowing for it to be implemented at the Sub Macroblock Prediction and Mode Decision stages of a block-based encoder. The model is based upon data which is gathered from the first three frames of the Foreman QCIF resolution video and subsequently tested on HD resolution against SATD to show the generalisation of the novel LHPSS model's applicability.

The results show the model is able to retain a level of IQ in IbBbBbBbP for similar or lower bit usage, though the time taken is high. For the IPPP prediction structure case, the bit usage is dramatically lowered at the expense of IQ, with time taken remaining high. The increase in time is related to LHPSS falling back on to SATD for those samples which SSIM or covariance values are out of scope of Pseudo-SSIM. While this novel approach shows a potential for bit-budget improvement, it can be refined with a more accurate model, which can address the IQ losses seen in IPPP. This can be addressed by extending the coverage of LHPSS to a wider range of SSIM and covariance from different video sources as discussed in [2]. Thus, produce a more accurate perceptual R-D model as mentioned in [3] as well as minimise the fall back to SATD, and so also address issues related to timing.

When the Intra frame image with the Mode Decision was shown, under LHPSS, homogeneous regions exhibited higher number of larger block sizes than SATD. This is encouraging, demonstrating that where the LHPSS model is successful in lowering the convex hull of the Rate-Distortion curve towards the origin. This work can be extended in the form of equation (3), so that Pseudo-SSIM adapts depending upon the perceptual nature of the block and the bit budget. Therefore, implementing a Local Perceptual Rate Control with a dynamically adapting perceptual distortion metric to complete the Perceptual Framework discussed in [2].

References

[1] H. Everett III, "Generalised Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources," *Operations Research*, vol. 11, no. 3, pp. 399 – 417, 1963.

[2] Y. G. Joshi, P. Shah, J. Loo, and S. Rahman, "Review of Standard Traditional Distortion Metrics and a need for Perceptual Distortion Metric at a (Sub) Macroblock Level," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting 2013*, June 2013.

[3] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 23–50, 1998.

[4] G. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, 1998.

[5] T. Richter, "SSIM as Global Quality Metric: A Differential Geometry View," in *Proc. Third Int Quality of Multimedia Experience (QoMEX) Workshop*, 2011, pp. 189–194.

[6] C.-H. Chou and Y.-C. Li, "A Perceptually Tuned Subband Image Coder Based on the Measure of Just-Noticeable-Distortion Profile," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 6, pp. 467–476, 1995.

[7] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[8] Y.-H. Huang, T.-S. Ou, and H. Chen, "Perceptual-Based Coding Mode Decision," in *IEEE International Symposium on Circuits and Systems (ISCAS), Proceedings of 2010*, 302010-june2 2010, pp. 393 –396.

[9] K. Sühring. H.264/AVC Reference Software JM. [Online]. Available: http://iphome.hhi.de/suehring/tml/

[10] S. G. Tan T.K. and W. T., "(VCEG-AJ10r1) Recommended Simulation Common Conditions for Coding Efficiency Experiments Revision 4." ITU-T SC16/Q6, 36th VCEG Meeting, San Diego, USA, 8th - 10th Oct., 2008, 2008.

$\mathbf{x}_{min}$	$\mathbf{x}_{max}$	$\mathbf{SSIM}_{cost}$ **Offset**	**Coeff.**	$\mathbf{Cov}_{cost}$ **ΔCov×x**	**ΔCov**
2	10	8	1 1/8	1/64	1/32
10	20	12	3/4	1/128	1/8
20	50	14	5/8	1/256	3/16
50	100	32	1/4	1/256	1/4
100	200	32	1/4	1/1024	5/16
200	600	32	1/4	1/512	1/2
600	800	−300	3/4	1/256	−1/2
800	900	−1,688	2 1/2	1/64	−10

TABLE II

4X4: ZONE1$_{Cov}$ (0≤COV<150) WHERE X=(1−SSIM)×1K

$\mathbf{x}_{min}$	$\mathbf{x}_{max}$	$\mathbf{SSIM}_{cost}$ **Offset**	**Coeff.**	$\mathbf{Cov}_{cost}$ **ΔCov×x**	**ΔCov**
2	10	12	3 5/8	1/256	1/16
10	20	28	1 15/16	1/256	1/16
20	50	42	1 1/4	1/256	1/16
50	100	58	15/16	1/512	1/8
100	200	80	11/16	1/512	1/8
200	600	88	5/8	1/512	1/8
600	800	−142	1	1/512	1/8

TABLE III

4X4: ZONE2$_{Cov}$ (150≤COV<300) WHERE X=(1−SSIM)×1K

$\mathbf{x}_{min}$	$\mathbf{x}_{max}$	$\mathbf{SSIM}_{cost}$ **Offset**	**Coeff.**	$\mathbf{Cov}_{cost}$ **ΔCov×x**	**ΔCov**
0	2	32	3 5/8	1/256	1/64
2	10	32	3 5/8	1/256	1/32
10	20	40	2 1/2	1/256	1/32
20	50	56	1 3/4	3/1024	1/32
50	100	86	1 1/8	1/1024	1/8
100	200	102	15/16	1/2048	3/16
200	600	108	7/8	1/1024	3/32
600	800	−120	1 1/4	1/128	−4

TABLE IV

4X4: ZONE3$_{Cov}$ (300≤COV<600) WHERE X=(1−SSIM)×1K

$\mathbf{x}_{min}$	$\mathbf{x}_{max}$	$\mathbf{SSIM}_{cost}$ **Offset**	**Coeff.**	$\mathbf{Cov}_{cost}$ **ΔCov×x**	**ΔCov**
0	2	36	5 5/8	1/64	1/256
2	10	36	5 5/8	1/256	3/128
10	20	56	3 3/8	1/512	5/128
20	50	80	2 1/4	1/1024	1/16
50	100	124	1 3/8	1/1024	3/64
100	200	160	1 1/8	1/2048	3/32
200	600	160	1 1/16	1/2048	3/32

TABLE V

4X4: ZONE4$_{Cov}$ (600≤COV<1350) WHERE X=(1−SSIM)×1K

$\mathbf{x}_{min}$	$\mathbf{x}_{max}$	$\mathbf{SSIM}_{cost}$ **Offset**	**Coeff.**	$\mathbf{Cov}_{cost}$ **ΔCov×x**	**ΔCov**
0	2	48	8 1/2	3/256	−1/64
2	10	48	8 1/2	1/512	1/64
10	20	84	4 3/4	1/1024	7/256
20	50	120	3	1/1024	1/32
50	100	176	1 7/8	1/1024	3/128
100	200	214	1 1/2	(ΔCV/8)×x/2048	7/64
200	250	284	1 1/4	(ΔCV/8)×x/2048	3/32

TABLE VI

4X4: ZONE5$_{Cov}$ (1350≤COV<3014) WHERE X=(1−SSIM)×1K

$\mathbf{x}_{min}$	$\mathbf{x}_{max}$	$\mathbf{SSIM}_{cost}$ **Offset**	**Coeff.**	$\mathbf{Cov}_{cost}$ **ΔCov×x**	**ΔCov**
0	2	66	12	3/512	−1/256
2	10	66	12	1/512	1/512
10	20	128	6 3/4	(ΔCV/16)×x/512	1/64
20	50	186	4 1/4	(ΔCV/4)×x/512	1/128
50	100	256	3	(ΔCV/4)×x/512	1/256

TABLE VII

4X4: ZONE6$_{Cov}$ (3014≤COV<8000) WHERE X=(1−SSIM)×1K

$\mathbf{x}_{min}$	$\mathbf{x}_{max}$	$\mathbf{SSIM}_{cost}$ **Offset**	**Coeff.**	$\mathbf{Cov}_{cost}$ **ΔCov×x**	**ΔCov**
2	10	42	3 1/4	1/64	5/8
10	50	46	2 1/4	1/64	3/4
50	100	96	1 1/4	1/64	1
100	600	128	1	1/128	3/2
600	750	−350	1 3/4	1/64	−3 1/4
750	900	−2,375	4 1/2	1/128	2
900	975	−22,000	26	1/128	1

TABLE VIII

8X8: ZONE1$_{Cov}$ (0≤COV<150) WHERE X=(1−SSIM)×1K

$\mathbf{x}_{min}$	$\mathbf{x}_{max}$	$\mathbf{SSIM}_{cost}$ **Offset**	**Coeff.**	$\mathbf{Cov}_{cost}$ **ΔCov×x**	**ΔCov**
2	10	150	5 1/2	1/16	1/4
10	50	150	5 1/2	1/64	1/8
50	100	192	4 1/2	1/512	1
100	600	256	2 1/2	1/256	1
600	900	−1,024	4 1/2	1/128	−1

TABLE IX

8X8: ZONE2$_{Cov}$ (150≤COV<300) WHERE X=(1−SSIM)×1K

$\mathbf{x}_{min}$	$\mathbf{x}_{max}$	$\mathbf{SSIM}_{cost}$ **Offset**	**Coeff.**	$\mathbf{Cov}_{cost}$ **ΔCov×x**	**ΔCov**
2	10	100	9	1/32	3/16
10	20	100	9	1/32	1/8
20	50	100	9	0	5/8
50	100	312	4	1/128	3/8
100	500	448	3	1/256	1/2
500	600	448	3	1/128	−3/2
600	750	−600	5	1/256	1/2

TABLE X

8X8: ZONE3$_{Cov}$ (300≤COV<600) WHERE X=(1−SSIM)×1K

$\mathbf{x}_{min}$	$\mathbf{x}_{max}$	$\mathbf{SSIM}_{cost}$ **Offset**	**Coeff.**	$\mathbf{Cov}_{cost}$ **ΔCov×x**	**ΔCov**
0	10	168	16 1/2	1/64	1/16
10	20	224	13 1/2	1/128	1/8
20	50	352	7	1/128	1/8
50	100	448	6	1/1024	3/8
100	550	552	4 3/4	1/512	1/4

TABLE XI

8X8: ZONE4$_{Cov}$ (600≤COV<1350) WHERE X=(1−SSIM)×1K

$\mathbf{x}_{min}$	$\mathbf{x}_{max}$	$\mathbf{SSIM}_{cost}$ **Offset**	**Coeff.**	$\mathbf{Cov}_{cost}$ **ΔCov×x**	**ΔCov**
0	10	245	30	1/128	1/32
10	15	368	17	1/128	1/64
15	20	368	17	1/256	1/16
20	50	464	12	1/256	1/16
50	100	464	12	−1/1024	5/16
100	275	848	6	1/256	0

TABLE XII

8X8: ZONE5$_{Cov}$ (1350≤COV<3014) WHERE X=(1−SSIM)×1K

$\mathbf{x}_{min}$	$\mathbf{x}_{max}$	$\mathbf{SSIM}_{cost}$ **Offset**	**Coeff.**	$\mathbf{Cov}_{cost}$ **ΔCov×x**	**ΔCov**
0	15	272	44	1/1024	1/128
15	50	512	20	1/512	1/64
50	200	952	10	1/1024	1/16

TABLE XIII

8X8: ZONE6$_{Cov}$ (3014≤COV<8000) WHERE X=(1−SSIM)×1K

Zone	**ΔCov Thresh**	**4x4: Coeff.**	**8x8:** $(Ax+B)\times\mu_\Delta+\mu_\Delta$
1	0	8	A: 1/16, B: 0
2	150	8	A: 1/16, B: 0
3	300	8	A: 7/128, B: -4
4	600	8	A: 3/512, B:16
5	1350	8	A: 5/512, B:2
6	3014	4	A: 5/512, B:-43

TABLE XIV

MEAN ABSOLUTE DIFFERENCE COST WHERE $\mu_\Delta = |\mu_{O-R}|$

Low complexity sub-block perceptual distortion assessment for mode decision and rate-control

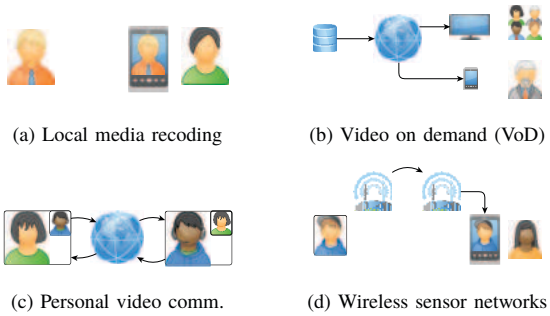
Y. G. Joshi, J. Loo, P. Shah, S. Rahman, and A. Tasiran
School of Science and Technology, Middlesex University, The Burroughs, Hendon, London, NW4 4BT, UK
Email: {y.joshi, j.loo, p.shah, s.rahman, a.tasiran}@mdx.ac.uk

***Abstract*---Video is being used in a variety of portable and low powered devices, via online/on-demand services, for personal communications or over heterogeneous/wireless sensor networks. Likewise, perceptual evaluation is being sought, to ensure video sequences maintain perceptual integrity. This raises a challenge, to bring high complexity perceptual algorithms into a low complexity environment. Existing perceptual solutions minimise the overall complexity by making the Lagrange multiplier (λ), the quantisation stage perceptually aware. These solutions are restricted to the block level using the original pixels, a model or previous encoded block, thus avoiding assessing individual sub-block candidates. Current perceptual algorithms like structured similarity (SSIM) uses statistical based calculations, and in order to be compatible with existing scores a further high complexity function is required for scaling. This paper presents a perceptual distortion and activity assessment that can operate at the sub-block for each candidate during the later stages of encoding, in mode-decision and in rate-control, without the need for statistical calculations nor the high complexity associated with scaling a perceptual algorithm. The paper will show how several perceptual techniques of SSIM luma function, just noticeable detection (JND) and a new proposed edge detection can be used to form a low complexity solution. Consequently, the proposed low perceptual assessment has additional timing increase of < +4 % for medium and low activity video sequences.**

I. Introduction

Video is increasingly generated by and delivered to low powered devices, which increases the mobility and accessibility of video. The traditional demands on the infrastructure of bandwidth remain, with the added factors of power consumption. Likewise, innovation is being sought for incorporating perceptual video coding (PVC) so that video encoding may retain perceptually significant features. As PVC is highly complex there is a challenge to make PVC within a low complexity envelope. The applications for such a solution are broad and can be placed into four major applications as shown in figure 1 of local storage, on-line/on-demand, personal video communications and wireless sensor networks. In fact, these applications can be drawn as storage critical or responsive critical. Figures 1a and 1b is where video is being stored or accessed for later period in time, which requires efficient use of available storage space. Figures 1c and 1d are where video depends on responsiveness and/or expects the network to incur dropped frames, meaning the video must not depend on any one frame.

In video coding, a frame is made from a mosaic of blocks, of different sizes, the choices of which reflect the balance between representing the content and regulating bandwidth.



(a) Local media recoding (b) Video on demand (VoD)

(c) Personal video comm. (d) Wireless sensor networks

Fig. 1: Applications for low complexity perceptual video coding

These square block sizes under high efficiency video coding (HEVC) can vary from 16x16 to is 64x64 and called the largest coding unit (LCU). The sub-block is any square size less than the LCU, represented as 2^n where n is a minimum of 2. Under rate-control, a fixed size of 8x8 is analysed for activity, while in mode-decision square sizes down to 4x4 are assessed for distortion. This is illustrated in figure 2 which leads to the choice of distortion metrics in the encoder. In the case of rate-control, a fixed 8x8 size variant of the Hadamard transform is applied, while for mode decision the block sizes can vary, leading to the use of a pixel based distortion assessment of sum of square errors (SSE).

Video encoding is about finding those combinations of sub-blocks or a single block for a given bit-rate constraint which can produce the minimum energy of distortion. For rate-control this means being able to adjust the quantisation to meet the bit-rate constraint. This is known as the Lagrange

(a) Mode decision: varying size, non-uniform distribution (b) Rate-control: fixed 8x8 size, uniform distribution

Fig. 2: 64x64: Sub-block size and distribution for mode decision (varying, non-uniform) and rate-control (fixed, uniform)

optimisation, where two opposing resources in this case rate and distortion is represented as a curve, here called an R-D curve, with the aim of finding the closest point to the origin on convex hull [1]. This is shown in equation (1)

$$J = D + \lambda \cdot R \tag{1}$$

where, J is energy of the (sub-)block, D is distortion based on the type of distortion metric used, λ is the optimal level of quantisation required to meet the constraint rate defined by R. Making video coding to be perceptually aware involves incorporating models based on the Human Visual System (HVS), so that features which are perceptually important are retained. This means identifying perceptually significant regions via a perceptual distortion assessment within mode decision so that smaller blocks are used where perceptually significant and larger blocks where perceptually homogeneous. Extending perceptual coding to activity assessment within rate-control will promote the bits budget distribution upon the partitioning or on the quantisation applied to the given block. Finally, as the distribution of bits and partitioning may change the overall frame measurements may remain the same, measuring these influences via the use of participants can be time-consuming. As such frame level measurements may be insufficient for where perceptual video coding affects the sub-block level, and more localised means are required to evaluate where and by what amount changes have occurred without the use of participants.

II. Literature Review

The human visual system (HVS) is regarded as a complex system, described as a series of stages. The initial stage has received much attention for several decades and models have been produced. However, only in recent years has there been interest on applying PVC on assessing the significance on the distortion than the content alone. Structured similarity (SSIM) is widely used as a perceptual measure of distortion, often cited as an alternative to peak-signal to noise-ratio (PSNR), and incorporated into block-based encoder as a form of PVC. The popular choice of SSIM as a perceptual assessment is because it is regarded as the least complex among its peers [2]. Unfortunately, when SSIM is compared to existing non-perceptual distortion metrics it is highly complex. There are solutions that incorporate SSIM and address the issue of complexity, they adapt when perceptual assessment occurs, from the frequent calls of mode-decision to the reduced calls of rate-control, which leads to a reduction in the overall complexity [3]. This means that at the rate-control stage a perceptual response model is produced by applying varying degrees of quantisations and calculating a curve of best fit. As such, a perceptual R-D curve guides the encoder, re-scaling λ as a perceptual (λ_p) based on non-perceptual distortion (D) during the mode decision stage. Furthermore, since SSIM scores are not compatible with existing scores, there is the need to scale within the same distortion range, often cited as the triangle inequality ($\triangleq$) problem with SSIM [4]. This additional stage of scaling can be even more complex than SSIM itself, which further compounds the issue of PVC being unattractive to low powered devices and leading to two major issues:

1) As a means to save complexity cost, λ as λ_p relies upon non-perceptual distortion assessment based upon pixel differences, avoiding the perceptual significance of the original pixel values.
2) The technique of λ_p means that perceptual distortion during rate-control activity is assessed ahead of mode decision at the frame or block level, thus avoiding perceptually assessing each mode decision candidate individually.

These limitations of perceptual quantisation mean that a perceptual distortion (D_p) solution which replaces D in equation (1) can potentially address these issues. Previous work into the behaviour of perceptual verses non-perceptual at sub-block level illustrates SSIM and existing non-perceptual distortion metrics, evaluate distortion differently [5]. Upon further investigation, covariance, a component of SSIM, was shown to correlate well with the perceptual model of Just Noticeable Distortion (JND), an insight that had not previously been presented [6]. This lead to the non-linear scaling of SSIM without the need of highly complex mathematical operations of logarithms or exponentials as provided by other SSIM scaled solutions. Despite this, the complexity of SSIM makes the solution unattractive at the sub-block level.

The approach by existing perceptual quantisation solutions have considered them from the block size, the LCU, ignoring the different sub-block structure as shown in figure 2. In terms of mode decision, the perceptual distortion assessment must support the different sub-block sizes as well as the various severity of distortion, leading to the use of SSE, pixel based distortion assessment [7]. With regards to rate-control, HEVC has replaced the mean absolute difference (MAD) distortion assessment with a variation of the Hadamard transform, a fixed 8x8 size without the DC value [8]. Consequently, any perceptual solutions will need to operate within the distortion metric space and the mode of operation of these existing non-perceptual measures to minimise complexity overhead. This means in mode decision a pixel-based perceptual solution and in rate-control a solution that can complement the existing Hadamard transform.

Typically, perceptual algorithms from imaging domain have been evaluated by participants, and in recently years image databases have used to evaluate different perceptual image quality assessments (IQAs) [2], [9]. This gives credence to incorporate perceptual IQA such as SSIM into tools to evaluate video encodings. However, the use of perceptual IQAs in the video domain on the decoded frame is limited. Video use of spatial and temporal techniques means that the search for minimum energy (J) as described in equation (1) is subject to managing bandwidth. This means that the encoder must dynamically adapt to the changing bit-target and result in different signalling choices. As these signalling choices are

governed by the bit allocation of rate-control and the search of J in mode decision this can affect the partitioning and level of quantisation. Visualising these signalling changes on the decoded frame can inform as to where the encoder is allocating bits and how broad or narrow the sub-block partitioning is to represent the content. Currently, commercial tools offer this feature, but non-commercial are limited to decoded video. Such a tool can be extended to support the development of new perceptual algorithms as a means to simulate or even verify its behaviour.

III. Research Design

PVC and low complexity can seem at odds with each other, as their is a risk to the robustness of any perceptual solution. Since perceptual IQA involves error normalisation, where differences are considered in their perceptual significance, finding methods to minimise complexity during normalisation is crucial. However, the use of perceptual IQA is not suitable for all conditions, it should be reserved for those sub-blocks where it is beneficial. This means that D_p score is a combination of traditional non-perceptual IQA and perceptual IQA, where perceptual IQA is added to those candidates which have a perceptually poor distortion. As such, there is a requirement to identify those sub-blocks which have a perceptually poor distortion without the complexity of perceptual normalising the distortion. It is proposed that to reduce the overall perceptual complexity a sample of the given sub-block candidate is assessed, from which further perceptual IQA is considered, as illustrated in figure 3.

Another issue raised is that the incompatibility of perceptual IQA and non-perceptual IQAs which leads to further complexity. This is partly due to the use of windowing by perceptual IQAs like SSIM to average changes within a given 8x8 window. Addressing this issue means moving away from fixed size windowing perceptual IQA, and into pixel based perceptual assessment. The benefit is that processor friendly techniques can be designed into perceptual normalisation to minimise complexity. Unfortunately, with pixel based perceptual IQA there is a risk of losing perceptual integrity and so a secondary low cost perceptual stage is required to increase perceptual robustness. Any secondary perceptual stage should consider adjacent pixels, the rate of change in perceptual normalisation, in the form of an edge-detection. Current edge-detections are complex and cannot fit within the smallest sub-blocks, hence a new edge-detection is required.

IV. Methodology

As stated above and shown in figure 3 normalising the entire sub-block can be expensive, instead sub-blocks should be sampled from which checks are made to identify if full perceptual assessment should occur. This can ensure additional processing is applied where necessary and limiting the overall complexity. The first task is to perceptually account the distortion of a sample sub-block so it is perceptually normalised than by pixel difference. Choosing those pixel locations that make up the sample sub-block is dependent upon the tests

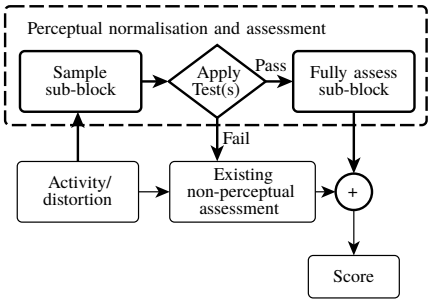


Fig. 3: Porposed perceptual normalisation and assessment workflow

to be applied. Here, a proposed test is where each side of the perceptually normalised version of the sub-block is taken (minus the corners) and subtracted from each other to find perceptually asymmetrical side labelled as PAS in equation (2).

$$PAS = (||T - B| - |L - R||) > Thresh \tag{2}$$

where T, B, L, R and $Thresh$ are top, bottom, left, right and threshold respectively. For mode-decision and rate-control the threshold values are 8 and 48 respectively, designed to reduce observations being processed by $\approx 50\%$, which are based upon analysis of 8x8 sub-blocks observations. Other specific tests to mode decision and rate-control are applied and are discussed below.

A. Mode Decision

Under mode decision the choices are between a combination of block sizes, a single block size or existing encoded block, whichever offers the least combined bit usage and distortion. At this stage of the encoder, each variation of the different block sizes represents the best prediction candidate, and so the distortion costs between these variations of block sizes may be minor. Whereas a traditional distortion metric seeks the minimum uniform cost, based on pixel difference, while a perceptual distortion assessment will consider the cost relative to the original pixel value. This means that the existing encoder workflow will need to be modified in order to pass the original pixel values at the last possible stage.

In the literature review, perceptual is calculated separately and not when existing distortion metrics are calculated. To reduce the overhead in perceptual processing, it should be integrated with non-perceptual, so the perceptual cost should be accumulated per pixel (like in non-perceptual) rather than averaged out across a block (like in SSIM). This means reducing the dependencies upon statistical calculations of mean, variance and covariance and using an in-line method to achieve similar forms of perceptual normalisation. While covariance is known to be a key perceptual component of SSIM, covariance in itself is still processor intensive. Instead, sum of square errors (SSE) should be adapted to support pixel based perceptually aware methods, in particular based upon the

calculation of covariance. Covariance involves both original (x) and reconstructed (y) pixels ($_i$), see equation (3).

$$\sigma_{x,y} = \frac{\sum (x_i \times y_i) - \Sigma x \times \mu_y}{n} \tag{3}$$

where $\sigma_{x,y}$ is covariance, μ is mean and n is the block size.

The sum product of original and reconstructed pixels can be rewritten as equation (4), where SSE is $(x-y)^2$.

$$\sum (x_i \times y_i) = \frac{\Sigma x_i^2 + \Sigma y_i^2 - SSE}{2} \tag{4}$$

It should be noted that in equation (4) the squared operation of original and reconstructed pixels can be represented by a look up table (LUT) to further reduce the complexity overhead. This allows covariance to be rewritten as equation (5).

$$\sigma_{x,y} = \frac{\frac{\Sigma x_i^2 + \Sigma y_i^2 - SSE}{2} - \Sigma x \times \mu_y}{n} \tag{5}$$

Unfortunately, equation (5) still involves one multiply and one divide for calculation of covariance, the divide must exist and cannot be substituted with a right shift. To reduce the complexity further while maintaining the perceptual properties, a new less intensive operation based on equation (4) is required.

1) Sum of square differences: During mode decision the variation between candidates can be very small, this means scores will need to be perceptually assessed at the pixel level. Combining a non-uniform cost with a uniform cost can risk destabilisation of the overall score, hence the perceptual cost should only represent a small proportion (within $\approx 10\%$). The proposed perceptual cost will be scaled down by a given factor depending upon its block size. This paper proposes SSE with Sum of Absolute Squared Differences (SASD), as shown in equation (6),

$$SASD = (\left|\left|x_i^2 - y_i^2\right| - SSE\right|)/2^8 \tag{6}$$

To ensure that SASD is used only where perceptual distortion activity is high, a threshold of $2^{(2n+3)}$ is applied, where n is $log_2(blockwidth)$, or as shown in table I. Also, since SASD is right-shifted by eight, the entire range of potential pixel costs can be stored in memory within a LUT with values between 0 and 255.

2) Edge detection: As described in section III a pixel based perceptual normalisation risks perceptual robustness and that another form of perceptual assessment is required. Here a new 2x2 edge detection is proposed, so that SASD is applicable where regions are textured.

Perceptual and non-perceptual are known to have similar correlation at either end of the scales, but differ in their response in between [10]. As such normalising distances in perceptual terms does not factor in the perceptual integrity of the block. To achieve this an edge detection should be applied on the perceptual normalised block. As most edge detections are large and cumbersome, a new proposed edge detection is presented in the form of a 2x2 sized edge detection as shown in equation (7),

$$Edge = (2 \cdot Centre) > (Top + Left) \tag{7}$$

where $Edge$ has a value of zero or one, and $Centre$, Top and $Left$ are pixel values. To keep the overall cost down four 2x2 edge detections will operate in a set pattern within every 4x4 block. As this edge detection is shaped as an `L', it can be orientated to test different pixels for perceptually significant boundary changes. The choice of this pattern of non-overlapping edge-detect provides coverage with minimum test points and can be referred to as 1/4 sub-block edge detection within mode decision. Finally, if sufficient edges are detected, then the block may add SASD cost to SSE as defined in table I.

B. Rate-control

During rate-control, block activity is used as a measure of how detailed the content is, and thus influence the number of bits to be allocated. This principle should be extended in perceptual terms, allowing bit allocation to be adjusted according to the perceptual activity. As rate-control activity employs a variant of Hadamard, any perceptual score should be applied in similar techniques to ensure consistent behaviour. This means that perceptual normalisation should be conducted under a similar pattern to Hadamard under rate-control, except that differences are perceptual and not pixel based. Since Hadamard will use pixel pairs of distances of 1, 2 and 4, the perceptual differences could be very large at times, making SASD unsuitable for perceptual activity under rate-control. A more advanced perceptual model is proposed which combines two perceptual models, SSIM luma function and just noticable difference (JND) background luminance masking. The SSIM luma function is described in equation (8), it is part of the trio of functions that eventually produce SSIM.

$$SSIM_l(x,y) = \frac{2\mu_x \times \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{8}$$

where C_1 is a constant based upon the maximum pixel range, [9]. While the JND background luminance masking is part of the JND perceptual measure and shown here in equation (9) [11], [12].

$$JND_l(x,y) = \begin{cases} 17 \times (1 - (\frac{bg(x,y)}{128})^{\frac{1}{2}}) + 3 & bg(x,y) \leq 127 \\ \frac{3}{128} \times (bg(x,y) - 127) + 3 & bg(x,y) > 127 \end{cases} \tag{9}$$

Threshold	4x4	8x8	16x16	32x32	64x64
Sum of squared differences	2^7	2^9	2^{11}	2^{13}	2^{15}
1/4 block edge detection	2^1	2^3	2^5	2^7	2^9

TABLE I: Mode decision - block size specific threshold values for both SASD and 1/4 block edge detection

Types of operations	Existing RC Had	Sample dom. side (stage 1)	Sub total	Cond. corner (stage 2)	Sub total	Diff. vs. RC Had	Proposed inside (stage 3)	Total	Diff. vs. RC Had	Altern. SSIM
Multiply, Divide	0	0	0	0	0	0	0	***0***	0	***208***
Addition, Subtract.	577	78	655	18	673	96	447	1120	543	329
Shifts	386	264	650	48	698	312	397	1095	709	0
Access LUT	0	72	72	12	84	84	108	192	192	0
Absolute	3	3	6	9	15	12	3	18	15	0
Branching	0	1	1	5	6	6	130	136	136	0

TABLE II: Complexity breakdown of proposed rate-control vs. non-perceptual and perceptual (SSIM without scaling)

where $(bg(x, y)$ is background luminance, in this case the higher of the two pixel pair values. These two perceptual models can be combined by first rearranged the SSIM luma function as $1-SSIM_l$, making it in-line with common perceptual principles. However, to consider this a perceptual cost, it should be scaled by the JND background luminance masking visibility threshold [11]. This is shown in equation (10)

$$Luma\,Cost(x, y) = (2^b - 1) \times (1 - SSIM_l)^{JND_l} \quad (10)$$

where b is bit-depth. Combining these two perceptual models in this way allows the SSIM luma function to be more non-linear due to the JND luma function, while the bit-depth range enables a pixel cost to be associated. This like SASD utilises a LUT to retrieve perceptual normalisation cost. Also the proposed luma cost uses the Hadamard transform to eliminate self-symmetry, but the cell distances of 1, 2 or 4, will be downscale by factors of $1/2$, $1/16$ and $1/64$ respectively. As an extension of the initial PAS test, a second stage of filtering is applied using these block sides which involves using the edge detection technique as discussed earlier, but on the respective corners. Since the perceptual normalised process involves Hadamard processing, the bottom right corner of the 8x8 block is always equal to zero and so is not tested.

The first two stages of the proposed perceptual rate-control activity cost work with the sub-block sample, 8x8 block's sides, and they act as an efficient means to eliminates false triggers, only permitting full assessment if it passes a series of thresholds. Table II shows the complexity breakdown of the proposed rate-control verses the existing non-perceptual and perceptual alternative SSIM. Since the proposed rate-control operates on top of the existing rate-control Hadamard (RC Had) function, the respective complexity are aggregated. As illustrated in figure 3, this means that the proposed algorithm is always more complex than the existing rate-control Hadamard function, but far less than SSIM based alternatives. Stages 1 and 2 reflect the two early termination points used to minimise perceptual overhead. While the total number of operations for performing the proposed algorithm is high, through the use of early termination points, stages of 1 and 2, full complexity cost are minimised. For sub-blocks which are no longer of interest the complexity cost can be counted at stages 1 and 2. Comparing the proposed perceptual approach to an existing perceptual technique of SSIM (based on the JM H.264/AVC [13], and assuming constants are pre-calculated) shows the high number of multiply and/or divide for the same 8x8 sub-block. Here, the complexity of SSIM is shown without considering the scaling that would need to occur. Therefore, the proposed perceptual rate-control works at the sub-block level of 8x8 with existing Hadamard based rate-control, offering early termination points whilst being processor-friendly.

C. Modified Decoder

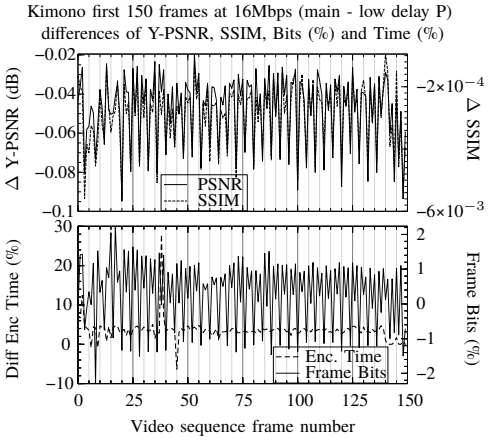
These proposed developments have been possible due to modelling based using empirical methods, but also by adapting the HEVC decoder as a tool to visually analyse their effect. Creating a visual perceptual assessment tool is motivated by the need to analyse how these new and existing techniques behave using real data. This becomes more crucial when visualising the partitioning and quantisation information from the signalling information. This is possible within the decoder by establishing a secondary stream reconstructed with the signalling information superimposed. To indicate the effect of these algorithms, a heatmap is produced, where blue is low and red is high. In certain cases a fixed sub-block assessment size of 8x8 is used to reflect the sub-block sampling approach used in the proposed algorithms.

V. Results and Discussion

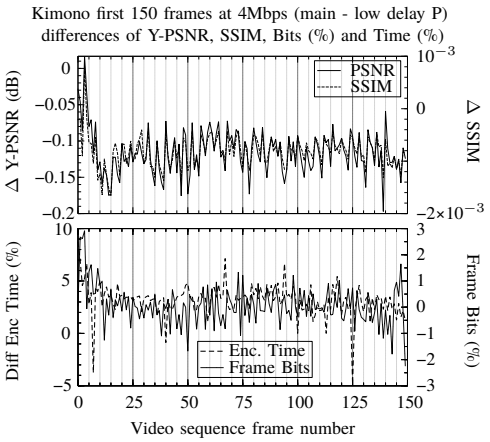
The low complexity perceptual solution proposed is designed to enable PVC in the scenarios illustrated in figure 1. These are aimed at capturing medium and low activity video, therefore, two high definition video sequences (of 1080p resolution) are used. For medium activity, BasketballDrive (50 fps), while for low activity Kimono (24fps), both encoded for 150 frames under the same HEVC main profile at 16Mbps, 4Mbps and 1Mbps. The use of BasketballDrive is to reflect the typical video being encoded of a set of large moving objects on a static background. Conversely, Kimono has a single person in the frame, similar to video calling scenarios, where background and foreground cam be more easily separated.

	Kimono Low Delay P			BasketballDrive Random Access		
Rate (Mbps)	Time (%)	Y-PSNR (dB)	SSIM	Time (%)	Y-PSNR (dB)	SSIM
1	3.06%	-0.1821	-0.0027	2.43%	-0.5298	-0.0110
4	3.12%	-0.1154	-0.0008	2.10%	-0.3353	-0.0047
16	3.51%	-0.0476	-0.0003	3.28%	-0.1045	-0.0012

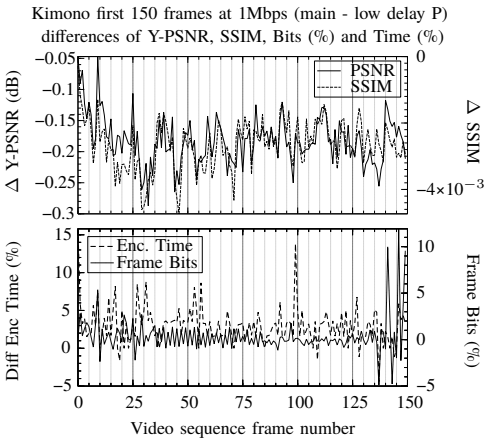
TABLE III: Overall difference by bit-rate for each video



(a) 16Mbps

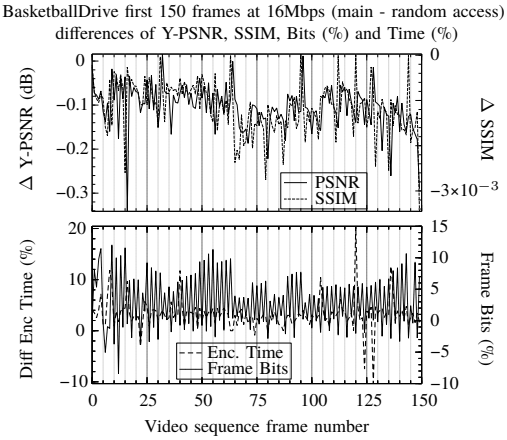


(b) 4Mbps

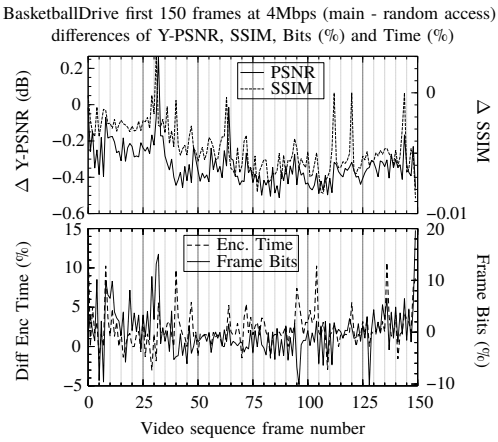


(c) 1Mbps

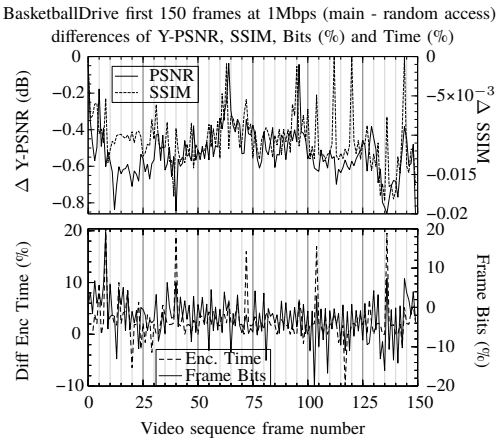
Fig. 4: Kimono first 150 frames



(a) 16Mbps



(b) 4Mbps



(c) 1Mbps

Fig. 5: BasketballDrive first 150 frames

As mentioned in the introduction the scenarios listed in figure 1 can be themed as either storage critical (pre-recorded video) or responsive critical (live interactive video) the two videos have been chosen for those themes. Here, BasketballDrive is selected for storage critical, as shown in figures 1a and 1b, while Kimono for responsive critical, figures 1c and 1d. To match these tests with their scenarios the configuration selected used are random access for storage critical themed scenarios, and low delay P for responsive critical scenarios. Random access is a hierarchical bi-predictive (B-frame) encoding with a group of picture (GOP) of 8 and intra refresh period of 32. Low delay p has a shorter GOP of 4 and uses hierarchical predictive (P-frame) where only the initial frame is an intra frame. In total 12 tests will be conducted altogether, 6 on the original unmodified encoder and 6 on the proposed low complexity perceptual encoder.

The proposed low complexity solution is built upon HEVC version 16 and thus tested against an unmodified HEVC version 16. The encoding timings have been gathered from a system running Ubuntu 15.04 with 7 Gb of RAM and running an Intel Core i7 processor at 2.67 GHz. To enable comparison the differences are reported, either as actual values or as percentages. Table III shows the overall difference of the proposed verses the original by bit-rate for each video sequence, the perceptual losses for both are minor, and the additional timing increases across the encoding sequence of the proposed solution is less than +4%. HEVC encoder produces logs enabling examination on a per frame basis and this is shown in figures 4 and 5, where Y-PSNR, encoding time and frame bits are presented alongside SSIM which is gathered via the modified decoder.

In the results for Kimono in figure 4 the loss in SSIM is insignificant (up to -0.0006, -0.002 and -0.004 for respective bit-rates) compared to the minor loss in PSNR (of -0.1, -0.2 and -0.3 dB respectively). Among the graphs, the changes in picture quality for both PSNR and SSIM tend to follow each other, with only the scale of losses differing. The lower SSIM losses may be attributed to non-perceptually sensitive regions, allowing the average PSNR to be lower, this suggests that the proposed encoder is redistributing bits on perceptual significant distortion. In terms of timing, they are fluctuating around the zero, although, there are instances when the timing differences are shown to dramatically increase or decrease. The extreme high and low in timings across the bit-rates do not cover the same period of frames, so these may be related to balancing content and bandwidth restrictions. In terms of the frame bit usage difference the dynamic is far less, within $\pm 3\%$, except for the least bit-rate of 1Mbps, this could be due to the profile of low delay P which stores additional prediction reference information.

For the results of BasketballDrive in figure 5, the use of random access and medium activity shows the loss in PSNR to be higher than in Kimono, starting from -0.3, then increasing to -0.6 and -0.9 as bit-rates decreases. The perceptual loss shows decreases of -0.003, -0.01 and -0.02 across the same bit-rates but linking set SSIM loss to PSNR loss is difficult due to nature of the video content and calculation of SSIM. Overall, the graphical changes of both PSNR and SSIM are similar, following each other but at times on lower bit-rates, SSIM peaks towards 0, when PSNR reports losses of -0.4 dB. This indicates that PVC can accept losses in PSNR without affecting the overall perceptual integrity of the video sequence. Looking at the frame encoding timings, there are significantly more peaks and troughs in random access compared to low delay P. These may be due to the high compression and intensive searching during motion estimation for B-frames, which is affected by the encoders ability to find J_{min}. For the proposed solution this affects the number of candidates to consider for full perceptual coding which can affect the final selection. As the algorithms are sub-block based, a visual simulation of their effects on an encoded frame may provide some insight of how the perceptual algorithm is behaving.

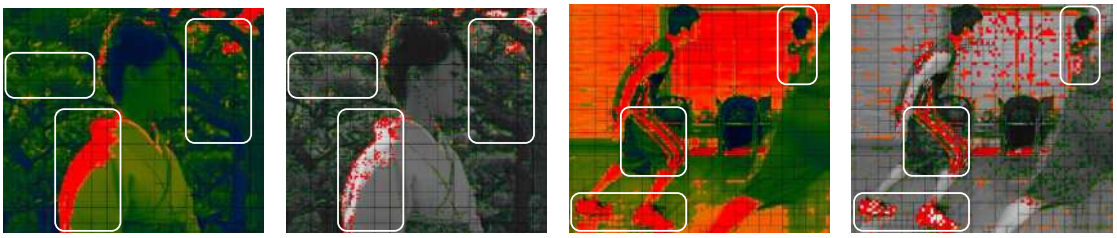
The modified decoder can assist to understand the proposed encoder's behaviour as it shows the effects of signalling in terms of sub-block partitioning, quantisation and by allowing the simulation of different distortion algorithms. Using the encoded bitstreams and extracting a frame from the middle of the sequence, in this case frame 77, the cropped heatmap images under different forms of assessment are shown in figures 6 to 8. Please note rate-control activity for both existing and proposed use the incoming original video sequence frame, so its results will be identical irrespective of bit-rate.

Figure 6 indicates how the proposed rate control which works in addition to the existing rate-control activity assessment places greater activity cost for high intensity regions with texture. This means that during mode decision, the influence of rate-control in bit-budget allocation and the perceptual distortion cost during RDO can lead to larger blocks in perceptually homogenous regions. Figures 7 and 8 show distortion measured in non-perceptual (SSE), perceptual (SSIM), proposed (SASD) and by quantisation (QP), where SSE and SSIM are heatmaps, while SASD and QP refer to heatmaps where triggered. This means that SASD reflects the research design as shown in figure 3 and for a low activity video sequence like Kimono even at 1Mbps there SSE and SSIM scores are low. QP reflects the level of quantisation set for the bits are allocated, otherwise if it is prediction only or block re-use it will remain greyscale.

The perceptually highlighted regions in figure 6b shows increased partitioning within the proposed encoder under figure 7. Similarly, in figure 8, the use of larger block sizes leads to very similar SSIM heatmaps. Overall, the proposed encoder use of larger blocks are on the boundaries between homogeneous and textured regions, suggesting some tolerance is provided under PVC.

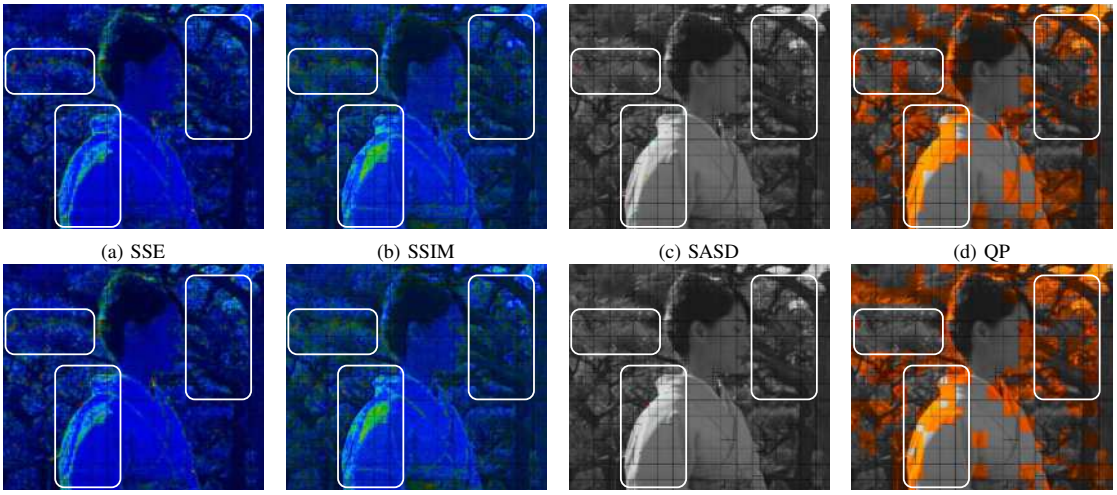
VI. Conclusion and future work

As more video services are accessed on portable devices, using heterogeneous and wireless sensor networks, the need for low complexity perceptual video coding increases. This paper presented, a low complexity perceptual solution capable for assessing candidates individually for both medium and low



(a) Existing (b) Proposed (c) Existing (d) Proposed

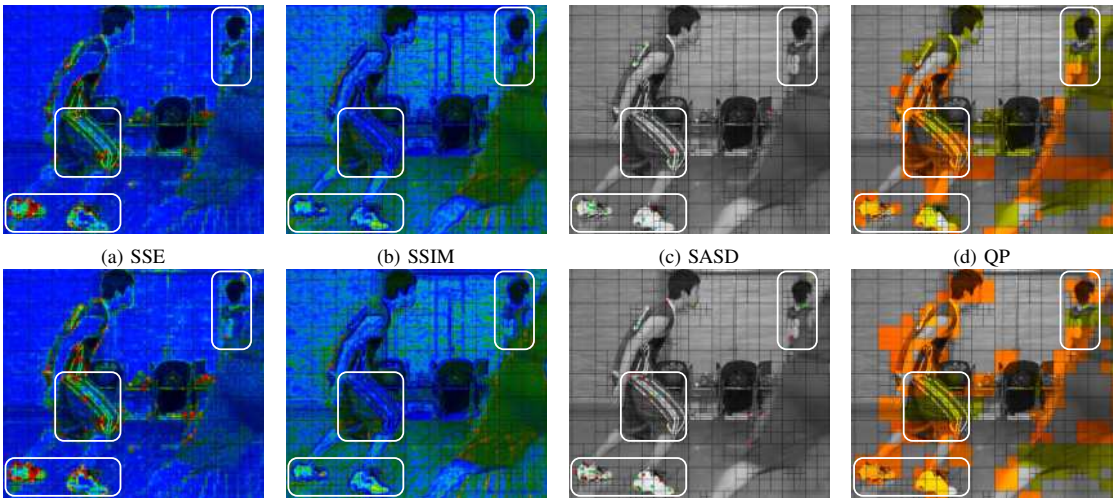
Fig. 6: Rate control activity on Kimono and BasketballDrive - blue is low, red is high



(a) SSE (b) SSIM (c) SASD (d) QP

(e) SSE (f) SSIM (g) SASD (h) QP

Fig. 7: Kimono at 1 Mbps, top row existing, bottom row proposed - blue is low, red is high



(a) SSE (b) SSIM (c) SASD (d) QP

(e) SSE (f) SSIM (g) SASD (h) QP

Fig. 8: BasketballDrive at 4 Mbps, top row existing, bottom row proposed - blue is low, red is high

activity video sequences. This can result in perceptually robust encodings for minimal increases in timing, < +4%. Other perceptual solutions for rate-control and mode-decision adapt existing perceptual algorithms with high levels of complexity and not for access each candidate. This paper presents a perceptual solution without the need for statistical calculations of SSIM or the high complexity operations used to scale SSIM into existing distortion metric space. As such the proposed solution can evaluation each mode-decision and rate-control candidate individually while maintaining a low complexity overhead. This should be extended to the prediction stage, the initial stage where sub-block assessment occurs, however issues of complexity tackled in this paper require greater sensitivity during prediction.

References

[1] A. Ortega and K. Ramchandran. ``Rate-distortion methods for image and video compression". In: *IEEE Signal Processing Magazine* 15.6 (1998), pp. 23–50. ISSN: 1053-5888.

[2] Weisi Lin and C-C Jay Kuo. ``Perceptual Visual Quality Metrics: A Survey". In: *Journal of Visual Communication and Image Representation* 22.4 (2011), pp. 297–312.

[3] Yi-Hsin Huang, Tao-Sheng Ou, Po-Yen Su, et al. ``Perceptual Rate-Distortion Optimization Using Structural Similarity Index as Quality Metric". In: *IEEE Transactions on Circuits and Systems for Video Technology* 20.11 (2010), pp. 1614–1624. ISSN: 1051-8215.

[4] T. Richter. ``A Global Image Fidelity Metric: Visual Distance and its Properties". In: *Image Processing (ICIP), 2013 20th IEEE International Conference on*. 2013, pp. 369–373.

[5] Y.G. Joshi, P. Shah, J. Loo, et al. ``Review of Standard Traditional Distortion Metrics and a need for Perceptual Distortion Metric at a (Sub) Macroblock Level". In: *Broadband Multimedia Systems and Broadcasting (BMSB), 2013 IEEE International Symposium on*. 2013, pp. 1–6.

[6] Y.G. Joshi, J. Loo, P. Shah, et al. ``A novel low complexity Local Hybrid Pseudo-SSIM-SATD distortion metric towards perceptual rate control". In: *Broadband Multimedia Systems and Broadcasting (BMSB), 2013 IEEE International Symposium on*. 2013, pp. 1–6.

[7] Hong Ren Wu, Weisi Lin, and King Ngi Ngan. ``Rate-perceptual-distortion optimization (RpDO) based picture coding --- Issues and challenges". In: *Digital Signal Processing (DSP), 2014 19th International Conference on*. 2014, pp. 777–782.

[8] Jens-Rainer Ohm, Gary J. Sullivan, Benjamin Bross, et al. *High Efficiency Video Coding (HEVC)*. Joint Collaborative Team on Video Coding (JCT-VC). 2013.

[9] Zhou Wang, A. C. Bovik, H. R. Sheikh, et al. ``Image Quality Assessment: From Error Visibility to Structural Similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612.

[10] A. Horé and D. Ziou. ``Is there a Relationship between Peak-Signal-to-Noise Ratio and Structural Similarity index measure?" In: *IET Image Processing* 7.1 (2013), pp. 12–24.

[11] Chun-Hsien Chou and Yun-Chin Li. ``A Perceptually Tuned Subband Image Coder Based on the Measure of Just-Noticeable-Distortion Profile". In: *IEEE Transactions on Circuits and Systems for Video Technology* 5.6 (1995), pp. 467–476.

[12] Tung-Hsing Wu, Guan-Lin Wu, and Shao-Yi Chien. ``Bio-inspired Perceptual Video Encoding based on H.264/AVC". In: *Proc. IEEE Int. Symp. Circuits and Systems ISCAS 2009*. 2009, pp. 2826–2829.

[13] Karsten Sühring. *H.264/AVC Reference Software JM*. URL: http://iphome.hhi.de/suehring/tml/.

Low complexity in-loop prediction perceptual video coding for HEVC

Y. G. Joshi*, J. Loo*, P. Shah*, S. Rahman*, A. Tasiran*, and J. Cosmas†
*School of Science and Technology, Middlesex University, The Burroughs, Hendon, London, NW4 4BT, UK
Email: {y.joshi, j.loo, p.shah, s.rahman, a.tasiran}@mdx.ac.uk
†Multimedia and Broadcast Networks Group, Department of Electronic & Computer Engineering,
Brunel University, Uxbridge, Middlesex. UB8 3PH, United Kingdom
Email: {john.cosmas}@brunel.ac.uk

***Abstract*—This paper applies the concept of hybrid framework for perceptual video coding (PVC) during the 'in-loop' stages by extending it to the prediction stage. As low complexity environments of mobile phones and tablets are increasingly used to capture video, PVC is not occurring here due to the high complexity of perceptual algorithms. Being able to encode using PVC will enable distortion to be merited by non-linear perceptual means than by uniform cost. While ideally, existing perceptual assessments of Structural Similarity (SSIM) is used, it is not processor friendly. The hybrid framework involves applying an additional low complexity perceptual assessment on top of existing Sum of Absolute Differences (SAD) and Sum of Absolute Transform Differences (SATD) only where distortion is perceptually significant. Consequently, the results show an increase in timing of $< +4\%$ and $< +6\%$ for video encoded with low delay P and random access profiles respectively, which is complexity competitively to other PVC solutions. This also affects bit redistribution with large reductions in bits allocated to signalling, -5 to -25%, with increases in small, medium and large block sizes. Visually, the proposed encoder encourages larger blocks on perceptually homogeneous regions and more dynamic smaller block where boundaries for textures or activity is occurring. This work can be extended to allow for perceptual quantisation to enable bandwidth reduction while maintaining perceptual quality.**

I. Introduction

Increasingly video is encoded on portable devices and shared across the Internet. Conversely, video consumption represents over 75% of all Internet and over half of mobile-data traffic [1]. Similarly, there is a decline in traditional computers and a growth in portable devices, allowing greater computing mobility [2]. While advances in video codec design on being hardware friendly have allowed video based applications on devices with limited processing resources. Therefore, video encoding on phones, tablets and cameras have become commonplace in everyday environments.

Underlying all these means of extending video into new applications is the same principle of video encoding, to represent video content for a given bandwidth. Traditionally, for a hybrid block based encoder, changes are evaluated at the block level, within 'in-loop' stages of prediction and mode decision, evaluating to find the optimal sub-block candidates or combinations of sub-blocks respectively. This evaluation involves measuring distortion (D) for the level of quantisation (λ) applied to the bits used to represent the differences (R), as shown by Equation (1) [3].

$$J_{min} = D + \lambda \cdot R \tag{1}$$

Equation (1) can be represented as a convex hull curve, known as the rate-distortion (R-D) curve, and by adjusting λ the desired bit-rate or distortion level can be met [4]. This means that for higher bandwidths greater bit information is retained, which for a hybrid block based encoder largely means less quantisation and an increase in the use of smaller sub-blocks. As video content changes, the respective R-D curve would differ, causing a different response R-D curve as illustrated in Figure 1.

Part of the R-D evaluation is distortion assessment, it can influence which sub-block candidates or combinations of sub-blocks are selected. Distortion assessment in video coding standards, differences are treated with uniform cost, irrespective of content, which does not match the non-linear response of the Human Visual System (HVS) [5], [6]. The HVS is highly complex and not fully understood, however, through experiments, theories and models have been devised which provide some insight [7], [8]. This desire to understand the HVS has been fuelled in the image coding domain, leading to the several models of Contrast Sensitivity Function (CSF), Just Noticeable Distortion (JND) and a perceptual assessment of Structural Similarity (SSIM) [9], [10], [11], [12].

II. Background

The progression of video coding into low powered devices has been possible due to video coding standards integrating process friendly techniques, as seen with H.264/Advance Video Coding, (H.264/AVC). Unfortunately, perceptual models and assessments are not designed for such environments and led to solutions which offer overall low complexity across an encoded sequence but operate outside the native 'in-loop' stages.

Existing PVC solutions tend to use SSIM as perceptual means to assess distortion as is it both perceptually effective and least complex among its peers [13]. Yet, SSIM is complex relative to existing standard traditional distortion metrics (STDM) of Sum of Square Errors (SSE), Sum of

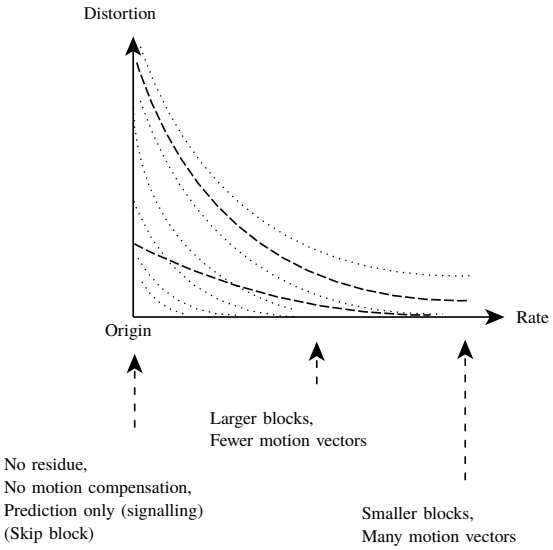


Fig. 1: Rate-Distortion curves

Absolute Difference (SAD) and Sum of Absolute Transform Difference (SATD). Consequently, existing PVC solutions are implemented out-side of the native front-end 'in-loop' stages of prediction or mode-decision as illustrated in figure 2.

Furthermore, SSIM is an index and does not provide the capabilities of a distortion metric since it does not conform to the triangle equality rule ($\triangleq$) [14]. This means that SSIM is scaled using non-linear means to be compatible with existing STDMs, which introduces further complexity, leading to PVC solutions which reside outside of the native sub-block level [15], [16]. These existing 'out-of-loop' PVC solutions operate by transforming non-perceptual scores to another value by mapping against a perceptual R-D curve as shown in figure 3.

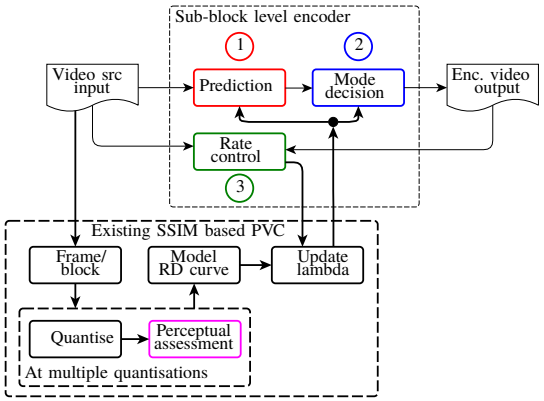


Fig. 2: Existing approach to perceptual video coding

This results in a single perceptual R-D model applicable to all mode decision choices, than the ideal of each prediction candidate or mode decision combination being evaluated on their own perceptual merits [17], [18], [19].

Ideally, distortion should be assessed perceptually, however, issues of complexity and compatibility hinder it adoption in limited processing environments. This has led investigations to understand how SSIM and STDM behave, labelling the relationship as a non-linear and geodesic-$\triangleq$ [14], [20]. When this relationship was explored at the 'in-loop' stage of prediction it revealed a shared space, revealing that perceptual assessment can vary depending upon video content [21]. Furthermore, an insight was uncovered that a component of SSIM, covariance was the key indicator to map SSIM to STDM, leading to a novel means to scale SSIM reusing existing values and low complexity techniques [22]. However, despite encouraging larger block usage, SSIM is complex relative to existing STDMs, which can justify pursuing an 'in-loop' PVC solution to influence the sub-block level choices.

To address the high complexity of perceptual assessment a framework was proposed, to apply perceptual assessment to candidates on a conditional basis, subject to whether significant perceptual distortion or activity [23]. This meant developing a new set of perceptual algorithms based upon earlier findings and existing perceptual models. Also, to keep computational demand low, pre-tests were created that sampled the sub-block candidates to determine whether perceptual assessment should occur. This meant an 'in-loop' PVC solution operated with $< +4\%$ increase in timing for medium and low activity videos, however, it lacked a solution for the most complexity sensitive stage of prediction. More recently, the complexity of previous perceptual PVC solutions is being discussed, and with JND on mode decision transform stage [24]. This is welcomed, yet it does not operate at the prediction stage which is highly complexity sensitive.

III. Methodology

This paper uses the hybrid framework as a means to produce a perceptual 'in-loop' solution, which means that a new algorithm, a series of tests and a set of thresholds must be presented [23]. The conditional framework demonstrated that

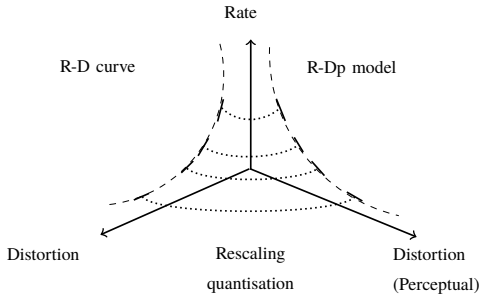


Fig. 3: Mapping R-D (STDM) curve against a R-Dp (perceptual) model to rescale quantisation

with an algorithm designed specifically for prediction the norm space is (ℓ_1) and with suitable pre-tests, both compatibility and complexity can be managed. In particular, this means that the decision on whether perceptual assessment should occur is governed by the threshold for the respective sub-block.

A. In-loop prediction algorithm design

Distortion assessment at the prediction occurs with SAD and Hadamard assessments, both of which are ℓ_1 norm space. This limits what type of perceptual choices are available for use, however, the same norm space is represented in rate-control where the hybrid framework was initially present [23]. Using the same process a new algorithm of additional pixel cost (APC) will be presented, based upon the SSIM Luma equation, as described in Equation (2),

$$SSIM_l(x,y) = \frac{2\mu_x \times \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{2}$$

where μ is mean and C_1 is a constant based upon the maximum pixel range. $C_1 = KL^2$, where K is 0.01 and L for 8 bit Luma is 255, [12]. To use SSIM Luma equation as a perceptual assessment, its scale must be extended to support the same norm-space and transformed to be tolerant of darker regions than lighter regions. This would invert the equation and match the desired norm space, however, as the rate of perceptual cost is linear, the equation should be accelerated with the perceptual model of JND using Equation (3), [11], [25].

$$JND(x,y) = \begin{cases} 17 \times (1 - (\frac{bg(x,y)}{128})^{\frac{1}{2}}) + 3 & bg(x,y) < 127 \\ \frac{3}{128} \times (bg(x,y) - 127) + 3 & bg(x,y) \geq 127 \end{cases} \tag{3}$$

where $bg(x,y)$ is the background luminance, in this case the higher of two pixel pair values. JND is a model based on the original frame only and has a non-linear response curve, having a greater tolerance for darker regions than medium to brighter regions As such, these two perceptual models of SSIM Luma and JND can be combined to form the proposed perceptual assessment. This is done by rearranging the SSIM Luma function as described above and making it in-line with common perceptual principles, labelling it as $1-SSIM_l$. Then in order to consider this as a perceptual cost, it should be scaled by the JND background luminance masking visibility threshold, producing Equation (4)

$$APC(x,y) = (2^b - 1) \times (1 - SSIM_l)^{max(JND_x, JND_y)} \tag{4}$$

where b is bit-depth. and $max(JND_x, JND_y)$ refers to using the maximum value for either the corresponding original or reconstructed pixel, which could vary depending upon the level of quantisation.

B. Proposed APC cross corner subtraction (ACCS)

From the initial paper pre-tests were used to evaluate whether perceptual assessment was required creating the perceptual pre-test technique of perceptual asymmetric side (PAS) [23]. Prediction involves evaluating different sub-block sizes, which under HEVC the largest coding unit (LCU) can be up to 64x64 and can include asymmetric sub-block sizes (width or height of 12, 24, 48). This means that perceptually assessing these varieties of sub-block sizes can represent significant proportion of processing time. Under these circumstances sub-blocks need to be evaluated faster with less perceptual accuracy. From this understanding, PAS was adapted for the sub-block corners, to form Equation (5)

$$ACCS = |(A_{TL} - B_{BR}) - (C_{TR} - D_{BL})| > ACCS_{Thresh} \tag{5}$$

where A to D denote the sub-block corners, T is top, B is bottom, R is right, L is left and $ACCS$ means APC cross calculation subtraction. ACCS, is where the respective diagonal corners are subtracted from each other based on their APC values. The equation was designed to be low complexity, with only a single absolute function being used.

The use of ACCS on the block corners is limited to four test points and was designed under an 8x8 sub-block. The choice of 8x8 is based findings that 8x8 is the optimal window value for SSIM [26]. Therefore, every whole multiple of 8x8 should undergo a further process of ACCS to ensure that perceptual cost is applied where it is suitable. This can be illustrated with Figure 4 which shows that the two stage process as outer for the sub-block corners and then as inner, for every 8x8 within.

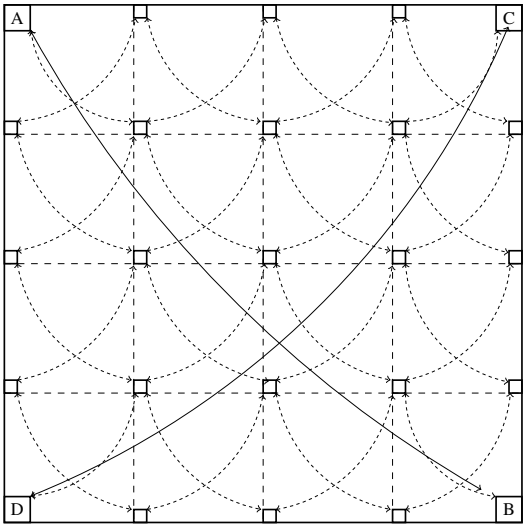


Fig. 4: APC cross calculation subtraction (ACCS) outer and inner (applicable to >8x8 sub-blocks)

C. Design for ACCS threshold in prediction

As mentioned earlier, the pre-test is threshold based, and the values must be derived to establish, however, as there are many combination of sub-blocks prediction there are multiple threshold values. Using raw Luma observations of an 8x8 sub-block the value of 128 was set, but later doubled to 256 when a high number of false triggers were observed during visual simulation. In addition, to extend this process to other sub-block sizes a non-linear scaling was formulated rounded to the nearest 8th in Equation (6).

$$ACCS_{Thresh} = (Int(log_{128}(2 \cdot blksize) \cdot 32)) \cdot 8 \quad (6)$$

where $blksize$ is the size of sub-block, width times height. Equation (6) was used to establish the thresholds for various combinations of sub-block sizes in HEVC as shown in Table I along with percentage equivalence to the initial 8x8. Equation (6) and Table I illustrate that a non-linear threshold discourages smaller sub-blocks as it has a relative lower threshold value compared to larger sub-blocks. Overall, this means the likelihood of applying additional perceptual cost is higher for smaller sub-blocks than for larger sub-blocks, which encourages the use of large sub-blocks.

IV. Results and discussion

The testing was conducted on an Intel Core i5-6600K system with 32GB of RAM using HEVC HM version 16.9. Eight video sequences were encoded, half under low delay P (LDP) and the other half under random access (RA) profile. Only the first half of each video sequence was encoded, 5 seconds worth as this has been deemed sufficient for testing [27]. Finally, each video sequence were encoded at 1, 2, 4, 8 and 16 Mbps. The encoder and decoder log files provided the Y-PSNR, timing and bit usage by block size results. While the SSIM scores were gathered using Video Quality Measurement Tool (VQMT) [28].

Blk Hght	Width (Threshold) 4	6	8	12	16	24	32	48	64
4	176	200	216	240	256	272	288	312	328
6	200	224	240	256	272	296	312	328	344
8	216	240	256	272	288	312	328	344	360
12	240	256	272	296	312	328	344	368	384
16	256	272	288	312	328	344	360	384	400
24	272	296	312	328	344	368	384	408	416
32	288	312	328	344	360	384	400	416	432
48	312	328	344	368	384	408	416	440	456
64	328	344	360	384	400	416	432	456	472

Blk Hght	Width (as %) 4	6	8	12	16	24	32	48	64
4	0.69	0.78	0.84	0.94	1.00	1.06	1.13	1.22	1.28
6	0.78	0.88	0.94	1.00	1.06	1.16	1.22	1.28	1.34
8	0.84	0.94	1.00	1.06	1.13	1.22	1.28	1.34	1.41
12	0.94	1.00	1.06	1.16	1.22	1.28	1.34	1.44	1.50
16	1.00	1.06	1.13	1.22	1.28	1.34	1.41	1.50	1.56
24	1.06	1.16	1.22	1.28	1.34	1.44	1.50	1.59	1.63
32	1.13	1.22	1.28	1.34	1.41	1.50	1.56	1.63	1.69
48	1.22	1.28	1.34	1.44	1.50	1.59	1.63	1.72	1.78
64	1.28	1.34	1.41	1.50	1.56	1.63	1.69	1.78	1.84

TABLE I: Non-linear threshold and percentage equivalent (with reference to 8x8) for APC cross corner subtraction (ACCS).

The results are shown in tables II and III indicating that the average PSNR losses are no more than 0.33 dB and 0.21 dB for LDP and RA respectively across all bit-rates. While the perceptual losses of 1-SSIM for both profiles are very minor < 0.003. Similarly, in terms of timing the increases for LDP and RA are $< +3\%$ and $< +6\%$ respectively including the upper standard deviation value. The changes to bit usage occurs in both LDP and RA, with significant reductions in signalling bits and increases in medium to large size blocks as shown in figure 5. The results shown are averaged across the five bit rates, however, to understand the range of differences the standard deviation (std dev) for bit changes are also shown. The std dev graphs show that the significant reduction in signalling is generally across all bit rates, with the exception of Riverbed which has the highest signalling reduction, which is probably due to the video content. In the video sequence Riverbed, the content is highly active and localised, making it very difficult to encode. Overall, the changes in bit usage is more dynamic in LDP than in RA, particularly for smaller block sizes.

Examining the decoded frames highlight which partitions are encoded with residual information. Examples comparing original and proposed for each profile are shown in Figures 6 and 7 for 1 and 16 Mbps respectively. The original frame images has partition information from the encoded bitstream is overlaid with colours of the rainbow representing each block sizes respectively, red (4x4), orange (8x8), yellow (16x16), green (32x32) and blue (64x64). In both Figures 6 and 7 white boxes have been superimposed to highlight particular features. For Figure 6 where there is low bit-rate, the proposed encoder is able allocate larger sub-blocks for homogeneous regions and smaller sub-block sizes closer to the video content. This is shown in Kimono where larger blocks are used on the background foliage, while in ParkScene smaller sub-blocks are allocated for the cyclists legs which are moving. While in Figure 7 where bandwidth is higher, the block redistribution is more dynamic. The proposed encoder identifies homogeneous regions where larger blocks or even skip blocks may be applied. This allows more smaller partitioning or sub-block sizes to be used where perceptually significant texture is present.

In the initial hybrid framework, only two video sources were used, one for RA and another in LDP [23]. This means that comparing these results with those presented in this paper is

Mbps	ΔY-PSNR	Δ1-SSIM	ΔTime	ΔTime Std Dev
1	-0.330	-0.0021	2.55%	0.74%
2	-0.221	-0.0013	2.61%	0.75%
4	-0.123	-0.0040	2.12%	0.89%
8	-0.069	-0.0005	1.79%	0.54%
16	-0.049	-0.0004	1.80%	0.28%

TABLE II: Average changes by bit-rate for low delay p

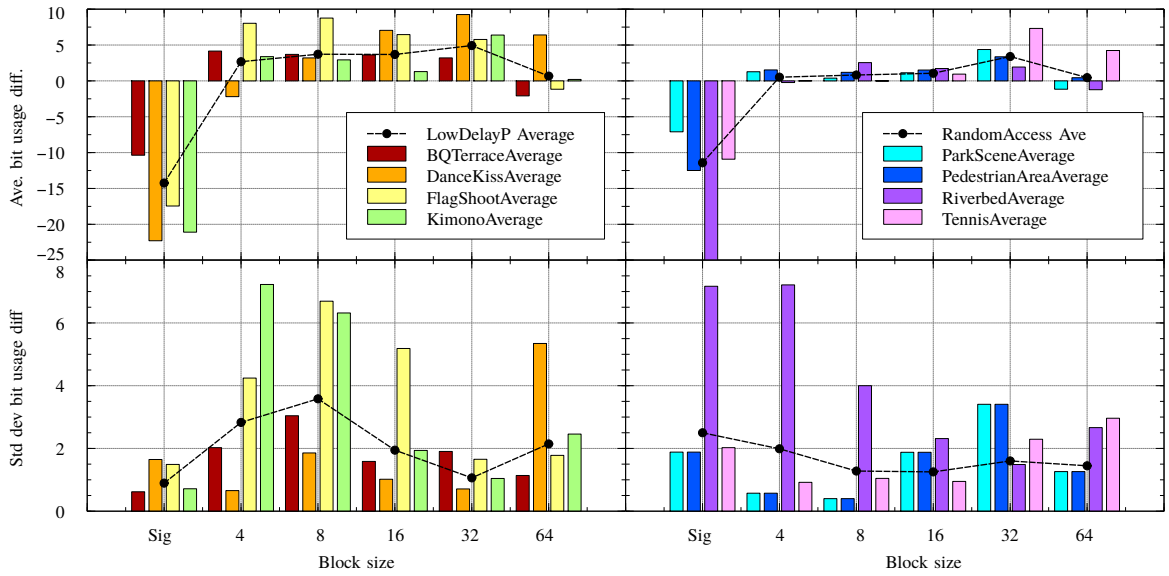
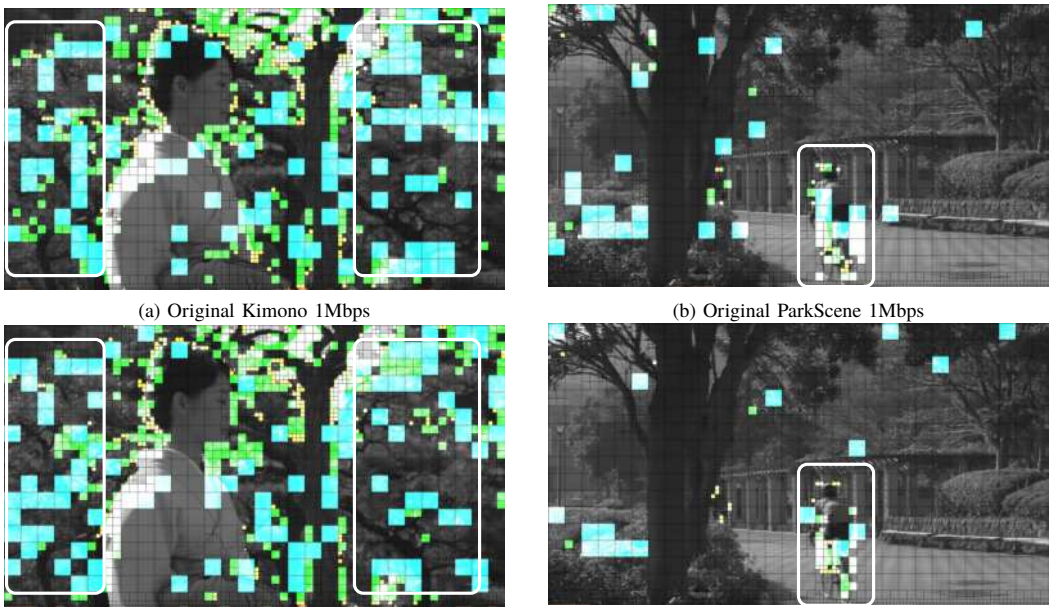


Fig. 5: Percent bit diff for low delay P and random access average (ave) and standard deviation (std dev) across bit-rates 1Mbps to 16Mbps by block size



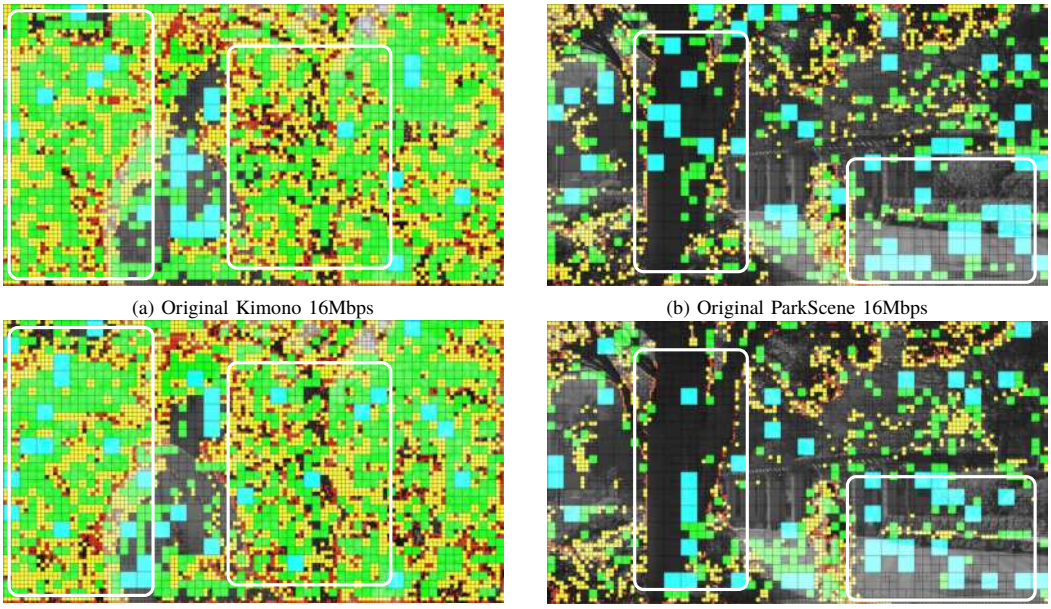
(a) Original Kimono 1Mbps

(b) Original ParkScene 1Mbps

(c) Proposed Kimono 1Mbps

(d) Proposed ParkScene 1Mbps

Fig. 6: Frame 77 for Kimono (low delay P) and ParkScene (random access) encoded sequences at 1 Mbps with highlight partition sizes



(a) Original Kimono 16Mbps

(b) Original ParkScene 16Mbps

(c) Proposed Kimono 16Mbps

(d) Proposed ParkScene 16Mbps

Fig. 7: Frame 77 for Kimono (low delay P) and ParkScene (random access) encoded sequences at 16 Mbps with highlight partition sizes

Mbps	ΔY-PSNR	Δ1-SSIM	ΔTime	ΔTime Std Dev
1	-0.205	-0.0026	4.95%	0.60%
2	-0.138	-0.0015	4.87%	0.92%
4	-0.086	-0.0008	4.64%	1.35%
8	-0.055	-0.0006	4.41%	1.17%
16	-0.035	0.0009	4.19%	1.07%

TABLE III: Average changes by bit-rate for random access

difficult. However, it is possible to say that by extending the hybrid framework to support prediction the timing increases from $< +4\%$ to $< +6\%$. In terms of PSNR and 1-SSIM losses, for LDP they are the same, while for RA they have been are reduced. This is especially true for 1-SSIM where at the lowest and medium bit-rate of 1 and 4 Mbps, the 1-SSIM losses are 1/4 of that observed previously, making the overall 1-SSIM losses virtually zero. Comparing the proposed solution with other 'out-of-loop' solutions, they offer bandwidth reductions through perceptual quantisation [15]. However, their high peak complexity makes them unsuitable for low powered applications. Where peak complexity is being addressed, the timing is still three times that shown here in this paper [24]. Therefore, the 'in-loop' solution presented here is complexity competitive to existing PVC solutions.

V. Conclusion and future work

Video encoding is increasing occurring on low powered devices, where the available processing is limited such as mobile phones and tablets. This presents a constrained complexity envelope for video coding to operate within. The proposed in-loop prediction PVC offers a low complexity means to encode video, suitable for the low powered devices. Having a in-loop PVC has shown it would allocate larger block sizes to perceptually homogeneous regions, thus allowing more dynamic partitioning on perceptually significant regions. In turn, this work can be extended from bit redistribution to bandwidth reduction with perceptual quantisation, making it more attractive solution to adopt.

References

[1] Cisco. *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014–2019*. Cisco. Feb. 2015. URL: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html.

[2] World Bank. *World Development Report 2016: Digital Dividends*. Tech. rep. World Bank, Jan. 2016.

[3] Hugh Everett III. "Generalised Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources". In: *Operations Research* 11.3 (1963), pp. 399–417. ISSN: 0030364X.

[4] A. Ortega and K. Ramchandran. "Rate-distortion methods for image and video compression". In: *IEEE Signal Processing Magazine* 15.6 (1998), pp. 23–50. ISSN: 1053-5888.

[5] H.R Wu and K.R Rao, eds. *Digital Video Image Quality and Perceptual Coding*. CRC Press, 2005.

[6] Hong Ren Wu, Weisi Lin, and King Ngi Ngan. "Rate-perceptual-distortion optimization (RpDO) based picture coding — Issues and challenges". In: *Digital Signal Processing (DSP), 2014 19th International Conference on*. 2014, pp. 777–782.

[7] P. Le Callet and E. Niebur. "Visual Attention and Applications in Multimedia Technologies". In: *Proceedings of the IEEE* 101.9 (2013), pp. 2058–2067.

[8] Mathieu Carnec, Patrick Le Callet, and Dominique Barba. "Objective Quality Assessment of Color Images based on a Generic Perceptual Reduced Reference". In: *Signal Processing: Image Communication* 23.4 (2008), pp. 239–256. ISSN: 0923-5965.

[9] J. Mannos and D.J. Sakrison. "The effects of a visual fidelity criterion of the encoding of images". In: *Information Theory, IEEE Transactions on* 20.4 (July 1974), pp. 525–536. ISSN: 0018-9448.

[10] J. Yogeshwar and R. J. Mammone. "A New Perceptual Model for Video Sequence Encoding". In: *Proc. Conf. th Int Pattern Recognition*. 1990, pp. 188–193.

[11] Chun-Hsien Chou and Yun-Chin Li. "A Perceptually Tuned Subband Image Coder Based on the Measure of Just-Noticeable-Distortion Profile". In: *IEEE Transactions on Circuits and Systems for Video Technology* 5.6 (1995), pp. 467–476.

[12] Zhou Wang, A. C. Bovik, H. R. Sheikh, et al. "Image Quality Assessment: From Error Visibility to Structural Similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612.

[13] Weisi Lin and C.-C. Jay Kuo. "Perceptual Visual Quality Metrics: A Survey". In: *Journal of Visual Communication and Image Representation* 22.4 (2011), pp. 297–312. ISSN: 1047-3203.

[14] T. Richter. "SSIM as Global Quality Metric: A Differential Geometry View". In: *Proc. Third Int Quality of Multimedia Experience (QoMEX) Workshop*. 2011, pp. 189–194.

[15] Yi-Hsin Huang, Tao-Sheng Ou, Po-Yen Su, et al. "Perceptual Rate-Distortion Optimization Using Structural Similarity Index as Quality Metric". In: *IEEE Transactions on Circuits and Systems for Video Technology* 20.11 (Nov. 2010), pp. 1614–1624. ISSN: 1051-8215.

[16] Wei Dai, O.C. Au, Wenjing Zhu, et al. "SSIM-based rate-distortion optimization in H.264". In: *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. 2014, pp. 7343–7347.

[17] Damon M Chandler. "Seven Challenges in Image Quality Assessment: Past, Present, and Future Research". In: *ISRN Signal Processing* 2013 (Nov. 2013).

[18] Po-Yen Su, Chieh-Kai Kao, Tsung-Yau Huang, et al. "Adopting Perceptual Quality Metrics in Video Encoders: Progress and Critiques". In: *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*. 2012, pp. 73–78.

[19] F. Zhang and D. Bull. "A Perception-based Hybrid Model for Video Quality Assessment". In: *IEEE Transactions on Circuits and Systems for Video Technology* PP.99 (2015), p. 1. ISSN: 1051-8215.

[20] A. Horé and and D. Ziou. "Image Quality Metrics: PSNR vs. SSIM". In: *20th International Conference on Pattern Recognition (ICPR), 2010*. Aug. 2010, pp. 2366–2369.

[21] Y.G. Joshi, P. Shah, J. Loo, et al. "Review of Standard Traditional Distortion Metrics and a need for Perceptual Distortion Metric at a (Sub) Macroblock Level". In: *Broadband Multimedia Systems and Broadcasting (BMSB), 2013 IEEE International Symposium on*. 2013, pp. 1–6.

[22] Y.G. Joshi, J. Loo, P. Shah, et al. "A novel low complexity Local Hybrid Pseudo-SSIM-SATD distortion metric towards perceptual rate control". In: *Broadband Multimedia Systems and Broadcasting (BMSB), 2013 IEEE International Symposium on*. 2013, pp. 1–6.

[23] Y.G. Joshi, J. Loo, P. Shah, et al. "Low complexity sub-block perceptual distortion assessment for mode decision and rate-control". In: *Broadband Multimedia Systems and Broadcasting (BMSB), 2015 IEEE International Symposium on*. 2015, pp. 1–9.

[24] Jaeil Kim, Sung-Ho Bae, and Munchurl Kim. "An HEVC-Compliant Perceptual Video Coding Scheme Based on JND Models for Variable Block-Sized Transform Kernels". In: *IEEE Transactions on Circuits and Systems for Video Technology* 25.11 (2015), pp. 1786–1800.

[25] Tung-Hsing Wu, Guan-Lin Wu, and Shao-Yi Chien. "Bio-inspired Perceptual Video Encoding based on H.264/AVC". In: *Proc. IEEE Int. Symp. Circuits and Systems ISCAS 2009*. 2009, pp. 2826–2829.

[26] A.C. Brooks, XiaoNan Zhao, and T.N. Pappas. "Structural Similarity Quality Metrics in a Coding Context: Exploring the Space of Realistic Distortions". In: *IEEE Transactions on Image Processing* 17.8 (Aug. 2008), pp. 1261–1273. ISSN: 1057-7149.

[27] F. Mercer Moss, K. Wang, F. Zhang, et al. "On the Optimal Presentation Duration for Subjective Video Quality Assessment". In: *IEEE Transactions on Circuits and Systems for Video Technology* (to be published). Early Access.

[28] VQMT: Tool. *VQMT: Video Quality Measurement Tool | MMSPG*. 2016. URL: http://mmspg.epfl.ch/vqmt.

A life is made up of a great number of small incidents and a small number of great ones.

Roald Dahl

Going Solo

People think that stories are shaped by people. In fact it's the other way around.

Terry Pratchett

Iorek: "You tricked Iofur Raknison?"
Lyra: "Yes. I made him agree that he'd fight you instead of just killing you straight off like an outcast, and the winner would be king of the bears. I had to do that, because—"
Iorek: "Belacqua? No. You are Lyra Silvertongue."
— Iorek Byrnison awarding Lyra the surname Silvertongue

Phillip Pullman

Northern Lights (The Golden Compass),
His Dark Materials Trilogy